Master's thesis
Master's Programme in Atmospheric Sciences
Physics/Aerosol Physics

Simple proxy for continental concentration of accumulation mode particles using reanalysis data

Aino Ovaska

August 9, 2021

Supervisor(s): Assistant Professor Pauli Paasonen

Examiner(s): Academician, Professor Markku Kulmala

University of Helsinki
Faculty of Science

PL 64 (Gustaf Hällströmin katu 2a), 00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Atmospheric Sciences Physics/Aerosol Physics

Tekijä — Författare — Author

Aino Ovaska

Työn nimi — Arbetets titel — Title

Simple proxy for continental concentration of accumulation mode particles using reanalysis data

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	August 9, 2021	65

Tiivistelmä — Referat — Abstract

Cloud condensation nuclei (CCN) participate in controlling the climate, and a better understading of their number concentrations is needed to constrain the current uncertainties in Earth's energy budget. However, estimating the global CCN concentrations is difficult using only localised in-situ measurements. To overcome this, different proxies and parametrisations for CCN have been developed. In this thesis, accumulation mode particles were used as a substitute for CCN, and continental proxy for number concentration of N_{100} was developed with CO and temperature as tracers for anthropogenic and biogenic emissions. The data utilised in the analysis contained N_{100} measurements from 22 sites from 5 different continents as well as CO and temperature from CAMS reanalysis dataset.

The thesis aimed to construct a global continental proxy. In addition to this, individual proxies for each site (the site proxy) and proxies trained with other sites' data (the site excluded proxy) were developed. The performance of these proxies was evaluated using a modified version of K-fold cross-validation, which allowed estimating the effect of dataset selection on the results. Additionally, time series, seasonal variation, and parameter distributions for developed proxies were analysed and findings compared against known characteristics of the sites.

Global proxy was developed, but no single set of parameters, that would achieve the best performance at all sites, was found. Therefore, two versions of global proxy were selected and their results analysed. For most of the sites, the site proxy performed better than the global proxy. Additionally, based on the analysis from the site excluded proxy, extrapolating the global proxy to new locations produced results with varying accuracy. Best results came from sites with low concentrations and occasional anthropogenic transport episodes. Additionally, some European rural sites performed well, whereas in mountainous sites the proxy struggled. Comparing the proxy to literature, it performed generally less well or similarly as proxies from other studies. Longer datasets and additional measurement sites could improve the proxy performance.

Avainsanat — Nyckelord — Keywords

CCN, N100, accumulation mode particles, proxy, CAMS

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

There are two main mechanisms through which aerosol particles affect the radiative budget of Earth and participate in controlling global temperatures [Boucher et al., 2013]. The first of them is associated directly with the radiative properties of particles, the second with interactions between clouds and aerosol particles. While both cause uncertainty to the current estimates of the global energy budget, especially aerosol-cloud interactions are not sufficiently quantified.

Clouds form when airmass becomes supersaturated with water vapour and the water starts to condense on existing aerosol particles called cloud condensation nuclei (CCN) producing cloud droplets [Boucher et al., 2013]. Clouds themselves have both cooling and warming effects on climate, as they enhance the planetary albedo but also contribute to the greenhouse effect by absorbing longwave radiation. These amount to cooling net effect. Particles impact cloud radiative forcing via number concentration of CCN available for producing cloud droplets. In a phenomenon called cloud albedo effect, a higher number concentration of CCN leads to the same liquid water content being distributed on a larger number of cloud droplets [Twomey, 1977]. This causes droplets to have smaller sizes and therefore larger total surface area, producing lighter clouds with higher albedo. In addition to the cloud albedo effect, smaller cloud droplets have been thought to increase cloud lifetimes as it takes longer for them to grow into sizes where they can precipitate [Boucher et al., 2013]. However, this effect is debated as small droplets are also more likely to evaporate [Small et al., 2009], which together with other microphysical changes might counteract the cloud lifetime increase [Toll et al., 2019].

Modelling the complex effects aerosol particles have on cloud albedo, lifetime and precipitation requires global quantification of particles able to activate into CCN [Rosenfeld et al., 2014]. This is not a simple task as CCN activation depends on multiple factors, including number size distribution and chemistry of the particles, but also meteorological conditions like supersaturation [Andreae and Rosenfeld, 2008, Paramonov et al., 2015, Boucher et al., 2013]. Direct CCN number concentrations can be measured using Cloud Condensation Nucleus Counters (CCNC), which optically measure the particles activated at different supersaturations [Paramonov et al., 2015]. While CCNC measurements are vital for validat-

ing models and understanding CCN behaviour at different locations, they are cost and labour intensive and have limited coverage [Schmale et al., 2018, Liu and Li, 2014, Rosenfeld et al., 2014]. Therefore, different parametrisations and proxies of CCN based on e.g. remote sensing and satellite observations [Rosenfeld et al., 2014, Liu and Li, 2014, Shen et al., 2019], bulk-chemical composition and number concentrations [Petters and Kreidenweis, 2007, Andreae and Rosenfeld, 2008], and other more widely available measurements [Nair and Yu, 2020] have been developed.

This thesis concentrates on developing a simple white-box proxy for estimating continental number concentrations of CCN-sized particles, which are defined here as accumulation mode particles. Accumulation mode describes particles that have diameters from 100 nm to around 2.5 μm [Seinfeld and Pandis, 2016]. This size range is characterised by weak removal processes which causes the particles to accumulate in the atmosphere. They correlate strongly with CCN and number concentration of particles with dry diameter larger than 100 nm (N_{100}) has been widely used as a proxy for CCN measurements. Since N_{100} measurements have better availability than direct CCN measurements, they are used also in this thesis.

Accumulation mode particles are produced either directly as primary particles or as secondary particles, which form from precursor gases in the atmosphere [Pierce and Adams, 2008, Seinfeld and Pandis, 2016]. Both primary and secondary particles can be anthropogenic or natural, as well as organic or inorganic [Boucher et al., 2013]. Based on modelling studies around 25-66% of global CCN at supersaturation 0.2% are from anthropogenic origin. At the same time research shows that natural biogenic emissions from ecosystems and the subsequent formation of secondary organic aerosol (SOA) also play a key role in significantly increasing CCN concentrations [Shrivastava et al., 2017, Paasonen et al., 2013, Riipinen et al., 2011]. Therefore, a successful N_{100} proxy must be able to capture both anthropogenic and biogenic sources. In this thesis, the proxy is constructed using air temperature (T) and carbon-monoxide (CO). [Paasonen et al., 2013] found that at higher temperatures N_{100} increases with increasing temperature due to ecosystem activity, whereas during cooler periods CCN-sized particles follow anthropogenic emissions. Based on this, temperature is selected to represent biogenic sources. CO, on the other hand, is produced during incomplete combustion, and therefore CO concentrations often correlate with accumulation mode particles formed in fossil fuel combustion and biomass burning [Zhou et al., 2020, Guyon et al., 2005], acting as a tracer for anthropogenic sources.

The proxy is developed using a dataset that contains daily averages of in-situ measurements of N_{100} from 22 stations located on 5 different continents. For CO and temperature Copernicus Atmosphere Monitoring Service (CAMS) reanalysis datasets produced by European Center for Medium-Range Weather Forecasts (ECMWF) [Inness et al., 2019a]

are used. The benefit of using reanalysis data for CO and temperature is that it has global coverage. Therefore, with a proxy developed with CAMS data it would be possible to predict N_{100} globally for the time period for which CAMS data is available.

This thesis gives the background information of how accumulation mode sized CCN particles are formed. The data and the method used for developing the continental N_{100} proxy are described. The aim is to produce a global continental proxy with one set of parameters, but also individual proxies for each site are created. The performances of the developed proxies are evaluated and compared against known characteristics of the sites. Finally, conclusions and recommendations for future work are presented.

2. Theory

2.1 Cloud-active aerosols

Only a subset of atmospheric aerosols can act as CCN. They are essential to cloud formation, as homogeneous formation of cloud droplets from pure water vapour would require relative humidities up to 800% [Andreae and Rosenfeld, 2008]. Therefore, in Earth's atmosphere, all cloud droplets form around cloud condensation nuclei or ice nuclei. In addition to available CCN, cloud formation requires also relative humidity above 100%, i.e. supersaturation, usually provided by updraft which brings the air-mass upward cooling it in the process [Rosenfeld et al., 2014]. Most of the atmosphere is CCN-limited, meaning that the actual number of activated cloud droplets at cloud base is defined by the number of CCN. Only in very polluted areas updraft limited conditions exist. Aerosol number concentrations and size distributions have large regional variation [Andreae and Rosenfeld, 2008] from very clean areas (few hundreds of CCN in cm^3 or below) to polluted areas (thousands of CCN in cm^3) [Rosenfeld et al., 2014, Schmale et al., 2018]. Also, the aerosol sources vary between locations and different sources have characteristic chemical compositions [Andreae and Rosenfeld, 2008, Boucher et al., 2013]. Whether a particle can activate into CCN at certain conditions can be estimated using Köhler theory.

2.1.1 CCN activation and Köhler theory

When using N_{100} as a proxy for CCN, it is assumed that particle's ability to act as CCN is mostly related to size. However, according to Köhler theory, the three main factors contributing to CCN activation are particle size together with chemical composition and ambient supersaturation [Andreae and Rosenfeld, 2008, McFiggans et al., 2006]. The minimum supersaturation required for the particle to become CCN is called critical supersaturation (S_c). It is defined by the curvature of the particle, called Kelvin effect, and equilibrium water vapor pressure of solution compared to pure water, called Raoult's effect, which relate to particle size and composition, respectively. According to Köhler theory, when supersaturation increases, a larger fraction of particles is activated. Big-

ger particles with good solubility are activated at lower supersaturations as they have more soluble molecules and smaller curvature, both factors that lower equilibrium vapour pressure and hinder evaporation [Andreae and Rosenfeld, 2008].

In reality, the effect of the chemical composition goes beyond solubility. Most particles are internally mixed, and additional factors, like presence of hydrophilic or hydrophobic, surface-active, or partially soluble substances, as well as surface films, wettability, and the shape of insoluble particle fraction, affect particle's ability to activate into CCN [Andreae and Rosenfeld, 2008, McFiggans et al., 2006]. This, together with the fact that aerosol composition has large variation, and mixing state is commonly not well know, makes it difficult to fully account for the effect chemical composition has on CCN activation. To overcome this, CCN-relevant chemistry is often described in terms of hygroscopicity, i.e. particle's ability to absorb moisture and different parameters based on this characteristic have been developed [Petters and Kreidenweis, 2007, Andreae and Rosenfeld, 2008].

However, research suggests that size is more important factor in predicting CCN activation [Dusek et al., 2006, Andreae and Rosenfeld, 2008, McFiggans et al., 2006]. This is possibly explained by the dependency between S_c and the number of soluble molecules, which is proportional to the third power of particle diameter and only linearly proportional to soluble mass fraction [Dusek et al., 2006]. Furthermore, due to atmospheric ageing, most particles contain at least some fraction of soluble compounds, causing size rather than composition to be a limiting factor in CCN activation [Andreae and Rosenfeld, 2008, Dusek et al., 2006].

Based on this, it is reasonable to assume that number concentrations of particles in relevant sizes can be used to approximate CCN number concentrations. However, this still leaves some variation in the size range, as activation diameter depends on supersaturation. According to literature, maximum cloud base supersaturation rarely exceeds 1% [Andreae and Rosenfeld, 2008], and the stratus clouds have maximum supersaturation around 0.2%, which corresponds to activation diameters larger than 80 - 100 nm [Pierce and Adams, 2008]. Other sources give similar estimations, with 50 - 400 nm as the CCN-relevant size range and dominant activation sizes from 50 nm to 200 nm [Dusek et al., 2006, McFiggans et al., 2006, Rosenfeld et al., 2014]. This makes most of the CCN accumulation mode particles, which is additionally supported by the observation that accumulation mode particles are more hygroscopic than smaller Aitken mode particles, due to a larger fraction of inorganic salts [McFiggans et al., 2006]. Therefore, N_{100} can be used as a proxy for CCN.

2.1.2 CCN sources

CCN-active accumulation mode particles have multiple different sources. Natural emissions can include, for example, sea spray from oceans, biogenic emissions from ecosystems, dust particles from soil and deserts as well as marine and volcanic sulphate emissions [Andreae and Rosenfeld, 2008, Boucher et al., 2013]. Major anthropogenic sources are fossil fuel combustion and biomass burning. However, in real atmosphere distinguishing between anthropogenic and natural aerosols is difficult, since most particles are mixtures of both [Andreae and Rosenfeld, 2008]. Additionally, some processes, like vegetation fires that contribute to biomass burning aerosol, are partially natural but enhanced by human interference. This further complicates making a clear distinction between natural and anthropogenic particles.

Another way of classifying particles is according to whether they are primary or secondary. Primary particles are interesting from CCN-perspective, because they directly add to particle number concentration, and are often produced in the correct size range [Andreae and Rosenfeld, 2008]. Additionally, new particle formation produces particles from the gas phase [Pierce and Adams, 2008]. In both cases, smaller particles go through secondary growth when atmospheric precursor gases condense on them [Pierce and Adams, 2008, Andreae and Rosenfeld, 2008]. Particle growth increases the fraction of particles large enought to act as CCN at given supersaturation. The relative importance of primary and secondary CCN depends on location.

When looking at primary CCN particles, the proxy developed in this thesis mostly accounts for anthropogenic primary particles. One reason for this is that many natural primary particle sources tend to produce larger coarse mode particles that have low number concentrations, and are not included in N_{100}. This is true for primary biogenic aerosol, which includes a large variety of particles such as detritus, pollen, spores, micro-organisms like bacteria, fungi, and viruses, and their fragments [Després et al., 2012, Boucher et al., 2013]. However, it should be noted that due to their size and solubility they can be important for CCN in locations with few other aerosol sources [Andreae and Rosenfeld, 2008]. Similarly, dust particles are typically coarse mode and only a small fraction is in accumulation mode [Boucher et al., 2013]. While large size contributes to high mass concentration, the number concentrations of dust particles are low and therefore they do not produce as many CCN [Andreae and Rosenfeld, 2008]. Additionally, since this thesis concentrates on a continental proxy, some primary particle species that have mainly marine sources are omitted even though they produce particles in accumulation mode. These include sea spray, which contains both sea salt and primary biogenic aerosol particles from marine ecosystems [Andreae and Rosenfeld, 2008, Boucher et al., 2013]. Also, sulphate aerosol has primary

emissions from oceans [Boucher et al., 2013], but in this thesis sulphate aerosol and its contribution to N_{100} is only considered through anthropogenic emissions which should correlate with CO.

The two anthropogenic sources this thesis concentrates on, fossil fuel combustion and biomass burning, produce both primary particles and gaseous emissions that contribute to particle growth. Biomass burning includes vegetation fires as well as domestic biofuel use and is one of the largest aerosol sources globally [Andreae and Rosenfeld, 2008, Reid et al., 2005]. Biomass burning particles have a varying chemical composition and consequently different hygroscopicities, but fresh particles contain mostly organic carbon, black carbon, and inorganic species [Reid et al., 2005]. Of these, inorganic compounds and around half of the organic fraction are soluble to water [Andreae and Rosenfeld, 2008]. Moreover, most of the mass and number concentration of biomass burning aerosol is produced in accumulation mode, allowing it to act as CCN [Andreae and Rosenfeld, 2008, Reid et al., 2005]. Atmospheric ageing further increases the likelihood of activation. Similarly, aerosol particles from fossil fuel combustion are also mostly carbonaceous. The CCN-active primary particles contain a combination of hydrophobic compounds, like black carbon and petroleum-based substances, along with soluble organic and inorganic compounds [Andreae and Rosenfeld, 2008]. The hygroscopicity of these particles depends strongly on the emission source, for example, pure hydrocarbon soot is quite hydrophobic, but soot from diesel engines already contains soluble substances like H_2SO_4. Based on modelling studies, primary anthropogenic particles from fossil fuel and biomass combustion account for a significant portion of CCN. [Spracklen et al., 2011] estimated that carbonaceous aerosol contributes 52-62% of surface-level CCN and according to [Merikanto et al., 2009] all primary emissions account for approximately 55% of global low-level CCN.

The gaseous precursors from fossil fuel combustion and biomass burning together with emissions from ecosystems participate in secondary aerosol formation. Secondary organic aerosol (SOA) forms in the atmosphere when volatile organic compounds (VOCs) oxidise, which lowers their volatility and increases the solubility, allowing them to condense on existing particles [Shrivastava et al., 2017]. VOCs from anthropogenic sources (AVOCs) include petroleum and combustion-related substances like alkanes, aromatic compounds, and polycyclic aromatic hydrocarbons [Andreae and Rosenfeld, 2008, Gentner et al., 2017]. Conversely, biogenic volatile organic compounds (BVOCs) cover organic gases emitted by terrestrial vegetation, for example, isoprene and terpenes [Shrivastava et al., 2017]. Globally, BVOCs dominate over AVOCs [Shrivastava et al., 2017] and contribute significantly to SOA [Shrivastava et al., 2017, Riipinen et al., 2011, Paasonen et al., 2013]. However, in addition to VOCs, other predominantly anthropogenic emissions like SO_2, NO_x, sulfate, nitrate, and ammonium enhance SOA formation by, for example, increasing oxidation and condensation rates

[Shrivastava et al., 2017, Andreae and Rosenfeld, 2008]. These also contribute to SOA hygroscopicity increasing the fraction of soluble compounds. The overall effect of SOA on CCN is significant. [Paasonen et al., 2013] estimated that biogenic SOA formation produces around half of N_{100} particles in Europe.

3. Data

Data used in this thesis contains 22 measurement sites displayed in figure 3.1. Most of the stations are in Europe, with four in Northern Europe (ASP, HEL, HYY, VAR), nine in Central Europe (ABZ, BSL, HPB, KCE, KPZ, MLP, NEU, SCH, WAL), and two in Western Europe (MHD, VIE). From other continents, three stations are in North America (ALE, EGB, SGP), two in South Africa (BOT, MAR), and one both in South America (SAO) and in Asia (NAN). All other locations were considered continental except Mace Head, Ireland (MHD), which is in a remote coastal area, as well as Hohenpeissenberg (HPB) and Schauinsland (SCH) in Germany, which are mountainous sites partially above the planetary boundary layer. ABZ, HEL, MAR, NAN, and SAO are urban sites, Värriö (VAR) remote site and Alert (ALE) remote polar site while the rest of the locations are rural. Further details and references for each station are in table 3.1.

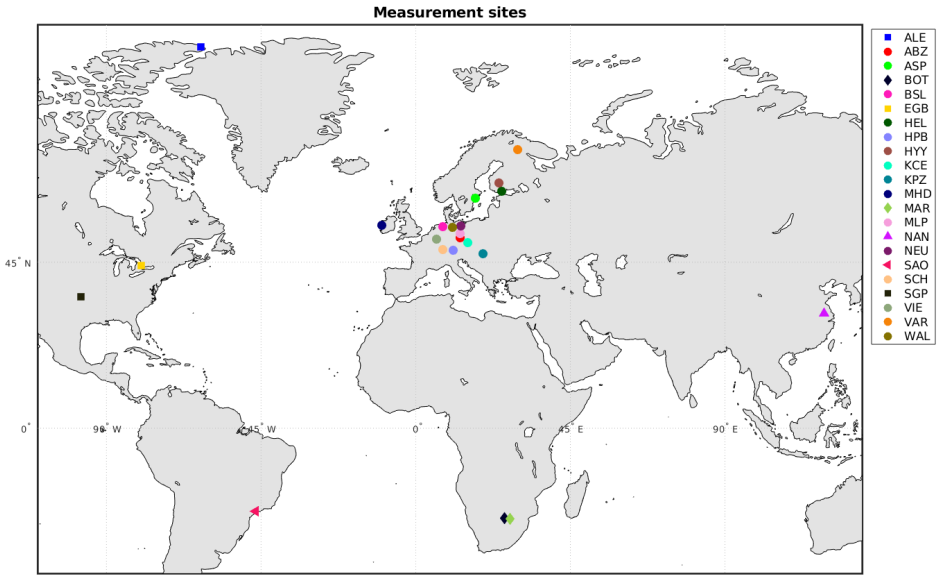

Figure 3.1: Map of all 22 measurement sites. Continents are distinguished with separate markers. Abbreviations and details for each site are shown in table 3.1.

Table 3.1: List of measurement sites, including abbreviation used in this thesis, site name and country, site description, coordinates, elevation in meters from ocean level and then the measurement instrument used for N_{100} measurements.

Abbreviation	Site and country	Site description	Coordinates	Elevation (m)	Instrument	Reference
ALE	Alert, Canada	Polar	82.492, -62.508	75	smps	[Nieminen et al., 2018]
ABZ	Annaberg-Buchholz, Germany	Urban background	50.57, 12.99	545	smps	[Birmili et al., 2016]
ASP	Aspverten, Sweden	Rural	58.8, 17.38	25	dmps	[Nieminen et al., 2018]
BOT	Botsalano, South Africa	Rural	-25.5, 25.8	1400	dmps	[Nieminen et al., 2018]
BSL	Bösel (Südoldenburg), Germany	Rural, regional background	53, 7.95	17	smps	[Birmili et al., 2016]
EGB	Egbert, Canada	Rural	44.2, -79.8	251	smps	[Nieminen et al., 2018]
HEL	Helsinki, Finland	Urban	60.2, 24.96	26	dmps	[Nieminen et al., 2018]
HPB	Hohenpeissenberg, Germany	Rural, hill	47.8, 11	988	smps	[Nieminen et al., 2018]
HYY	Hyytiälä, Finland	Rural	61.85, 24.29	181	dmps	[Nieminen et al., 2018]
KCE	Kosetice, Czech Republic	Rural, regional background	49.56, 15.08	534	smps	[Zíková and Zdímal, 2013]
KPZ	K-Puszta, Hungary	Rural, regional background	46.97, 19.55	125	dmps	[Nieminen et al., 2018]
MHD	Mace Head, Ireland	Remote, coastal	53.32, -9.88	10	smps	[Nieminen et al., 2018]
MAR	Marikana, South Africa	Urban	-25.7, 27.5	1170	dmps	[Nieminen et al., 2018]
MLP	Melpitz, Germany	Rural	51.53, 12.9	87	dmps	[Nieminen et al., 2018]
NAN	Nanjing, China	Urban	32.2, 118.9	25	dmps	[Nieminen et al., 2018]
NEU	Neuglobsow, Germany	Rural, regional background	53.14, 13.03	70	smps	[Birmili et al., 2016]
SAO	São Paulo, Brazil	Urban	-23.6, -46.6	750	dmps	[Nieminen et al., 2018]
SCH	Schauinsland, Germany	High-altitude, regional background	47.91, 7.91	1205	smps	[Birmili et al., 2016]
SGP	Southern Great Plains, Oklahoma, US	Rural	36.6, -97.5	300	dmps	[Nieminen et al., 2018]
VIE	Vielsalm, Belgium	Rural	50.3, 6	496	smps	[Fays et al., 2019]
VAR	Värriö, Finland	Remote	67.76, 29.61	390	dmps	[Nieminen et al., 2018]
WAL	Waldhof, Germany	Rural, regional background	52.8, 10.76	75	smps	[Birmili et al., 2016]

3.1 N_{100} measurements

N_{100} data was attained from in-situ measurements performed between 2003-2018. Depending on location, the measuring instrument used was either Differential Mobility Particle Sizer (DMPS) [Aalto et al., 2001] or Scanning Mobility Particle Sizer (SMPS) [Wiedensohler et al., 2012] (table 3.1). The measured size ranges vary for each site but typically cover particles from tens to several hundreds of nanometers [Nieminen et al., 2018, Birmili et al., 2016, Zíková and Zdímal, 2013]. N_{100} is calculated from particles between 100 nm and the upper limit of the measurement instrument. Most accumulation mode particles are observed within this size range.

For the analysis, daily medians were used. Figure 3.2 shows the time coverage for each station and missing data periods. The shortest datasets come from São Paulo, Brazil (SAO), with 283 days of data, and Egbert, Canada (EGB), with 356 days of data. The longest datasets are from Hyytiälä (HYY) and Värriö (VAR), in Finland, with measurements for the entire 2003-2018 time period.

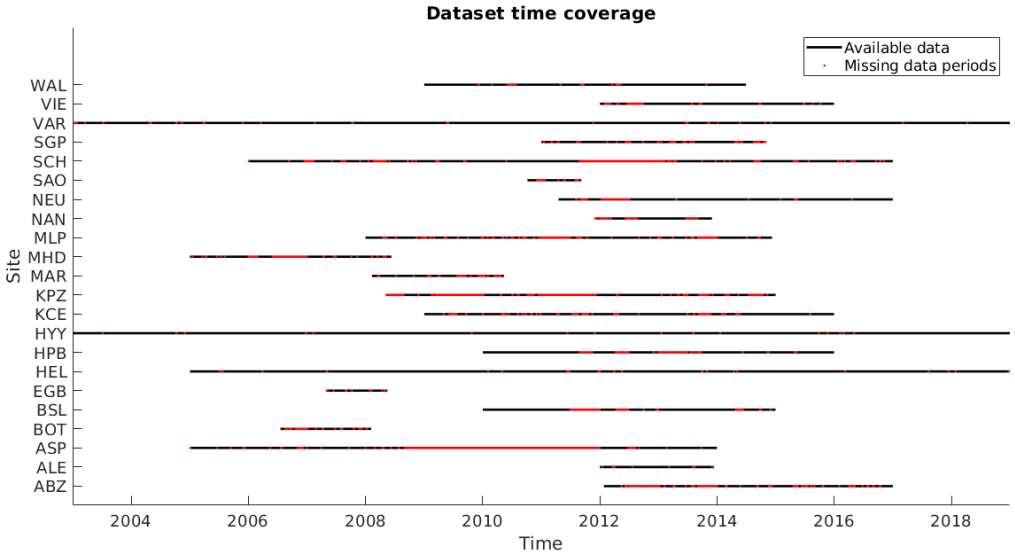

Figure 3.2: Time coverage and data availability of dataset for N_{100} at all sites.

3.2 CAMS reanalysis data

Reanalysis data is produced by assimilating together historical observations from different sources with past numerical weather broadcasts. Then a current numerical weather broadcast model is rerun using assimilated data as initial values. CAMS reanalysis data [Inness et al., 2019a] provides a global dataset for atmospheric composition using assimilated satellite trace gas data and ECMWF's Integrated Forecasting System model

[Inness et al., 2019b]. CO and temperature data were downloaded at model level 10 m above ground, using 0.75° grid around each measurement site with 3-hourly steps. Data was calculated into daily medians, which were used together with N_{100} to produce the proxy.

4. Methods

4.1 Proxy calculation

The developed proxy (eq. 4.1) has temperature dependant and CO dependant parts, described by empirical temperature parameter a(T) and CO parameter b_{ave}.

$$N_{100} = a(T) \cdot T + b_{ave} \cdot [CO] \tag{4.1}$$

The proxy was derived by first dividing N_{100} and CO data into temperature bins, which were defined so that the minimum bin width was 2°C and the maximum number of bins was 20. Next, linear fits for N_{100} against CO were calculated at each bin. As the proxy aims to estimate N_{100}, predicting the correct order of magnitude for the concentrations is more important than exact number concentration. Therefore, a fitting method that uses relative error was developed. First, a simple linear least squares fit was applied to the data points. After this, the method calculated the ratios of the observations and fitted line. The data points for which this ratio was outside two standard deviations were considered outliers. Also, the highest two percentiles of CO measurements were treated as outliers to remove the abnormally high CO peaks. Ignoring these outliers, a new fit was applied to the rest of the data points. This was done by using Matlab function *fminsearch* to minimise the sum of logarithmic ratios between data points and the fit

$$f = \sum_i \sqrt{(log(y_i/\hat{y}))^2}, \tag{4.2}$$

where y_i are the data points and $\hat{y}$ is the fit. The method repeated this process calculating new outliers for each iteration until the fit no longer changed, it stabilised between two possible fits, or it reached the selected maximum number of iterations. Minimising the logarithmic ratio i.e. the logarithmic distance between data points and the fit forces the line closer to lower data points while allowing larger absolute error for higher N_{100}. Figure 4.1 shows examples of the difference between the relative error method and linear least squares fit. The relative error method typically achieves better fit in clean conditions when both N_{100} and CO are low, though there are also bins where the difference between the two methods is minor.

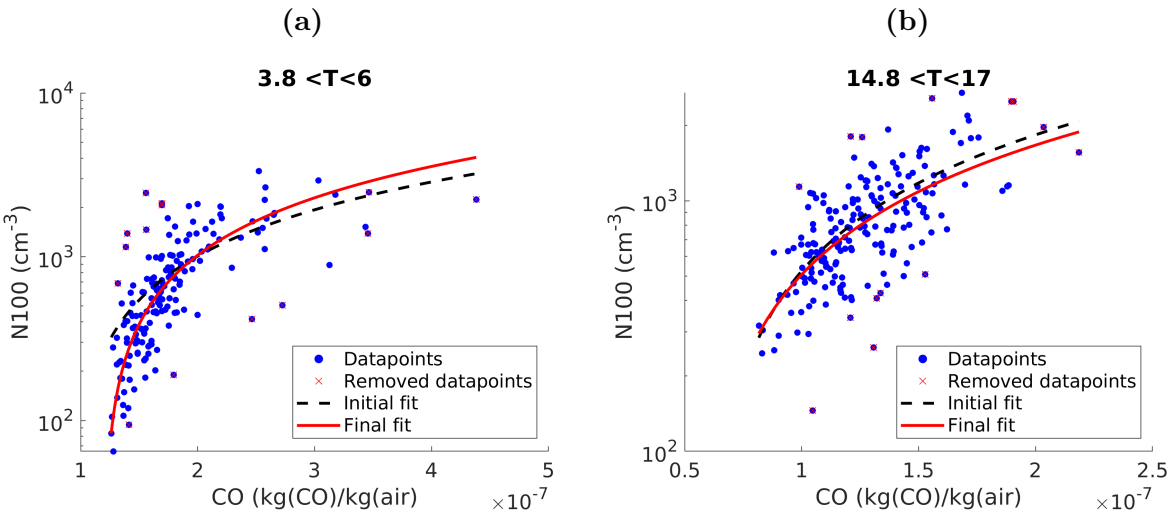

Figure 4.1: Figure shows difference between regular least square fit (initial fit) and the relative error method (final fit) a) for Neuglobsow (NEU) data in temperature bin 3.8<T<6, where relative error method fits better to small N_{100} values and b) NEU temperature bin 14.8<T<17 where the difference between fits is small. Removed data points indicate the outliers in final iteration.

Fits of N_{100} against CO concentration at different temperature bins can be seen in figure 4.2a, where Värriö is used as an example. Regardless of temperature, the slopes of the fits stay relatively constant. Based on this, the relation between N_{100} and CO concentration can be described using single parameter b_{ave}, which was calculated as the average of the slopes. Before averaging the slopes were additionally weighted with the negative logarithm of the p-value of the bin to ensure that the bins with the more robust correlation between N_{100} and CO contribute more to parameter b_{ave}. Bins where the p-value was greater than 0.05 were omitted from the analysis.

Based on the average slope b_{ave} new intercepts for each bin were calculated (figure 4.2b). For most sites, at lower temperatures the intercepts were constant and started to increase linearly towards higher temperatures. This is likely related to biogenic activity and the contribution of BVOCs to SOA at temperatures higher than 5°C, when the growing season starts. However, there were also sites where the intercepts were not constant below 5°C, but increased towards cold temperatures. This is possibly associated with boundary layer height decreasing during cold winter periods, which in turn enchances the concentrations. Despite this, for creating the proxy, constant temperatures below 5°C were used for all sites and the possible effect of boundary layer height was left out. To represent this, parameter a(T) was defined so that for temperatures below 5°C it used a constant p-value weighted average intercept. For temperatures above 5°C a similarly weighted linear fit was applied. The intersection of these lines was allowed to vary. For sites where there was less than two temperature bins above or below 5°C and, therefore,

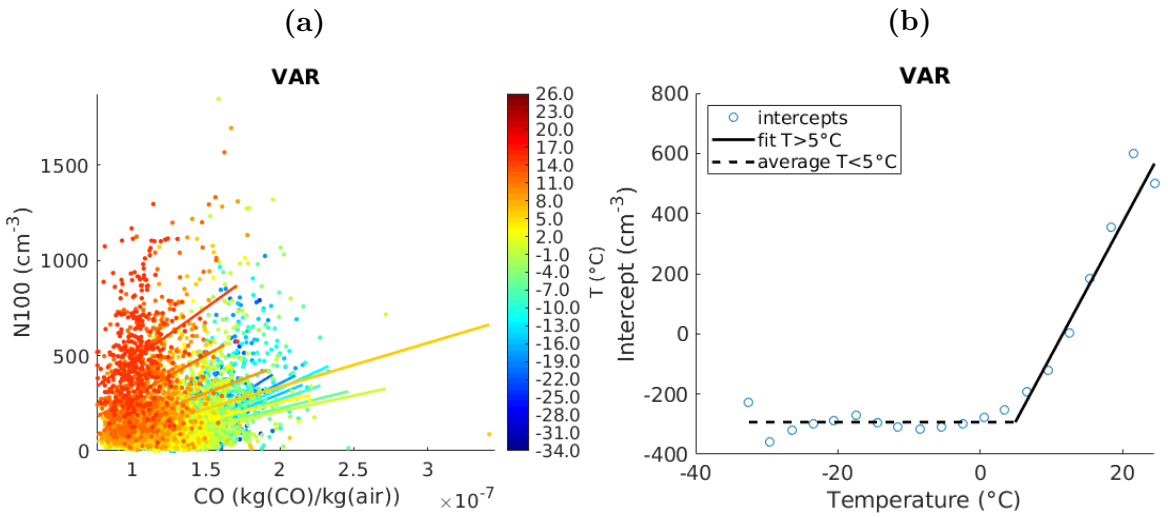

Figure 4.2: Panel a) has N_{100} as function of CO in Värriö (VAR). Colours show temperature bins with linear fits. Based on the slopes of the fits, CO dependant parameter b_{ave} was calculated. Panel b) shows intercepts, recalculated with constant slope b_{ave} from panel a), as function of temperature. Two fits were applied to intercepts for calculating temperature dependant parameter a(T): constant line for temperatures below 5°C and linear fit for higher temperatures.

one of the fits could not be calculated, the existing fit was extrapolated to all temperature bins.

Finally, the proxy was calculated by applying the parameters a and b_{ave} to eq. 4.1. Furthermore, the lower limit of N_{100} was set to 10 cm^{-3} to remove possible negative values. The concentration of 10 cm^{-3} corresponds to low values in Alert, which is the cleanest site in the dataset.

4.2 Proxy evaluation methods

For evaluating proxy performance, parameters were first calculated with a subset of data called training dataset. The rest of the data was used as a test set, where the N_{100} was predicted by applying test set temperature and CO concentration together with parameters from training set to eq. 4.1. Predicted N_{100} values were then compared to observed N_{100} from the test set.

When comparing observed and predicted N_{100}, implemented metrics were R^2 and root mean squared log error (RMSLE). R^2 is the square of Pearson's correlation between observed and predicted N_{100}, and it describes the fraction of the variance in observed N_{100} that is explained by the variance in proxy. RMSLE is an error measure that describes how much log-transformed predicted N_{100} values differ from log-transformed observed N_{100} (eq. 4.3). The benefit of using RMSLE instead of typically used RMSE is that it

penalises percentage error rather than absolute error. In other words, it gives same errors if two observed N_{100} values diverge from fitted value with same factor. For example, data points two times (200%) larger and 1/2 times (50%) smaller than fitted value would have same penalty. As a result, the fitting method penalises underestimation rather than overestimation. These attributes are similar to the relative error method used in calculating proxy parameters.

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(N_{100_{pred}}+1) - log(N_{100_{obs}}+1))^2} \tag{4.3}$$

Different training data selection leads to proxies that are tuned differently. The main aim of this thesis is to produce a global continental proxy, which is trained on all continental sites' data and yields one a(T) parameter and one b_{ave} parameter globally. However, to examine more in detail how the created model works, also proxies trained with a) only the site's own data and b) all other continental sites' data were created and their performance evaluated. These were named as "the site proxy" and "the site excluded proxy", respectively. Contrary to the global proxy, the site proxy and the site excluded proxy yield one set of parameters for each site.

One complication in selecting train and test sets was that the sites have varying number of data available (fig. 3.2). For example, training the global proxy with 16 years of data from Hyytiälä (HYY) and one year from Egbert (EGB) would bias the model. To overcome this, instead of using the entire dataset, the global proxy and the site excluded proxy were trained on one year of data from each site. This balanced the discrepancies between sites with shorter and longer datasets but still covered the entire seasonal cycle. However, when only one year of data is used in training, the varying conditions between years start to affect calculated parameters. For example, if one year in the dataset was exceptionally warm, the parameters trained with that year's data are not representative for other years. Also, the selection of test data affects the evaluation results. Therefore, in order to properly evaluate the performance of the different proxies and the effect train and test set selection have on the accuracy of the proxy, a version of K-fold cross-validation was utilised.

K-fold cross-validation splits the data into K-number of train and test sets typically so that test sets contain different data points [Bergmeir and Benítez, 2012]. Model is separately trained on each of these train sets and evaluated against the corresponding test set. The result gives parameters and evaluation metrics for each train and test set split. These can be used to estimate how much different data selections affect model performance.

K-fold cross-validation was directly applied when evaluating the site proxies. Here the proxy for each site was trained using only the site's own data. At each iteration,

one calendar year of data was used as a test set, and the proxy was trained with the rest of the data from the site. This was repeated until all years had been used as a test set. For each iteration, a separate proxy with its own parameters was created and R^2 and RMSLE values estimated. Here K depended on the number of years of data available. Additionally, some iterations with small test sets were manually excluded from the analysis if they had clearly deviating results. Small test set size was typically caused by years where the measurement period started or ended in the middle of the calendar year.

Site excluded proxies were trained with one year of data from all other continental sites. This required slightly adjusting the K-fold cross-validation method for train sets. Here, non-continental sites MHD, HPB, and SCH, together with the site in question, were excluded from training data. From each of the sites that were not excluded, one random year with at least 3/4 of the data was selected. For sites EGB and SAO, which have less than one year of data, all available data was always used. Because randomly selecting train data affects the resulting proxy, for each test set this was repeated five times. Test sets, in turn, were still handled for all sites by going through each available calendar year of data. Therefore, the number of iterations was five times K, where K was the number of calendar years. Of these, the iterations with small test sets that produced deviating results were manually excluded. For rest of the iterations, proxy was evaluated and R^2 and RMSLE values were collected.

Finally, the global proxy was evaluated utilising a similar method. Here, proxies for each site were generated using data from all continental sites. Again, for each calendar year used as a test set, five global proxies were generated by randomly sampling one year with at least 3/4 of data from all continental sites, and the R^2 and RMSLE values were saved. This resulted in five times K iterations for each site, from which iterations with small test sets were manually excluded.

For the site proxy, in addition to RMSLE and R^2, also the effect of seasonal variability on proxy performance together with time series were analysed. Seasonal analysis was done by dividing proxy into four three-month periods (December-February, March-May, June-August, and September-November) and compared to observed N_{100}. Time series were analysed as a general annual cycle of predicted and observed N_{100}. Additionally, predicted N_{100} time series from each train-test set split were compared to time series of observed N_{100}.

4.3 Final global proxy

After evaluating the proxies, the final global proxy was created. It was calculated by training the proxy with 100 different sets of training data, which contained one randomly

selected year of data, with at least 3/4 of data available, from each continental site. For SAO and EGB all available data was always used in training as they both have less than one year of data. With the trained global proxy parameters, N_{100} for each site was predicted using all that site's data as test set. The predicted N_{100} was then compared to the observed N_{100} from the site in question. This resulted in 100 sets of parameters, and 100 R^2 and RMSLE values for each site. R^2 and RMSLE values were analysed, and the parameter distributions were inspected including possible clustering of sites based on parameters. Two sets of global parameters were selected to visualise the predicted N_{100} against observed N_{100} and the results were connected to known characteristics of the sites. Finally, the ranges of observed N_{100} at each site were compared to ranges of predicted N_{100}, both with the entire dataset and seasonal division. Also median N_{100} values for predicted and observed data were compared.

5. Results

5.1 Proxy evaluation results

5.1.1 The site proxy

The results from evaluating the site proxy performance are displayed in figure 5.1. The boxplots show the variation in the evaluation metrics caused by different train and test set splits at each site. The number of different train-test selections is expressed with K. Some data selections, where the test set was small and the results clearly deviated from other results, were manually omitted from figure 5.1.

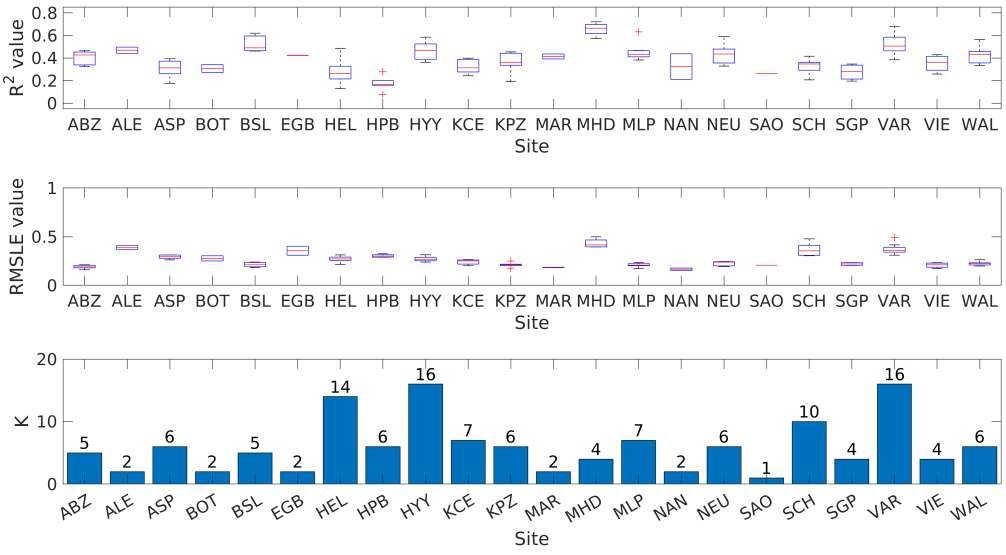

Figure 5.1: Figure shows how well the site proxy is able to predict N_{100} when trained with site's own data. The train and test sets were varied by leaving always one year out, using rest of the data to calculate the model parameters and then testing against the excluded year. Boxplots estimate the influence different data selection has on proxy performance. Red line shows median R^2 and RMSLE values, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and red crosses outliers. Additionally K value describes how many data points is included in boxplots.

Figure 5.1 clearly shows that both R^2 and RMSLE vary between the sites but also within each site. This internal variation is caused by the selection of train and test years. From physical perspective, this can be explained by different conditions. For

example, in ALE, which has only two years of data, the parameters calculated from 2013 data overestimate 2012 N_{100} and vice versa (figure 5.2). The problem arises from CO concentrations. While N_{100} between the years does not differ as much, CO concentrations are significantly lower during June-December 2013 compared to same time period in 2012. Therefore, the proxy parameters calculated with 2013 data predict stronger increase in N_{100} with increased CO concentrations, which leads to predicted N_{100} being overestimated for 2012. Conversely, proxy underestimates N_{100} for 2013 when using 2012 as training set, though the proxy lower limit at 10 cm^{-3} partially hides this. Without the lower limit, predicted values would be negative.

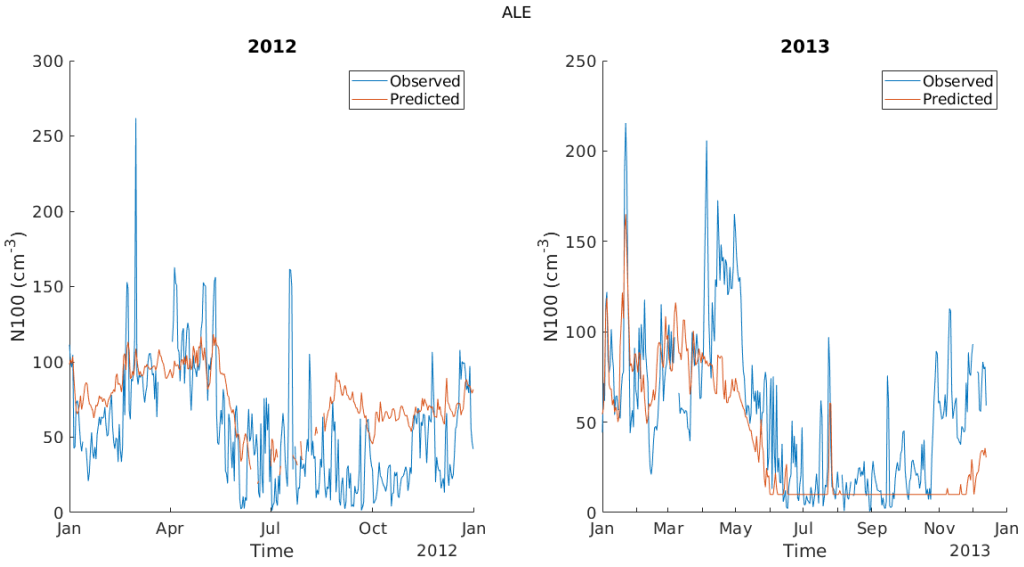

Figure 5.2: Time series of observed and predicted N_{100} for Alert, Canada (ALE). Proxy was trained with site's own data so that when predicting N_{100} for 2012, proxy parameters were calculated using 2013 data. Then N_{100} was predicted with 2012 temperature and CO using calculated parameters. Same procedure was repeated for 2013 proxy.

However, on top of physical features also statistical aspects need to be considered when analysing the effect of train-test set splits. Some of the internal variations between the years are caused by the different number of data points used for testing and training. Even though the test sets with poor data availability and related outlier results were excluded, it is impossible to avoid completely the differences in dataset sizes because of gaps in data coverage. This is more of a problem at some sites than others. For example, in Helsinki (HEL) the train sets contain always 4699-4715 data points and test sets 350-366 data points, so the variation is caused mainly by physical differences between years. In K-Puszta (KPZ), on the other hand, the train sets vary between 1110-1404 data points and test sets 68-362 datapoints, which is likely to explain at least some of the variation. However, the effect of the number of data points is not direct so that a small number of

data would automatically result in better or worse results. For example, if the test set covers only part of the year, also seasonality plays a role.

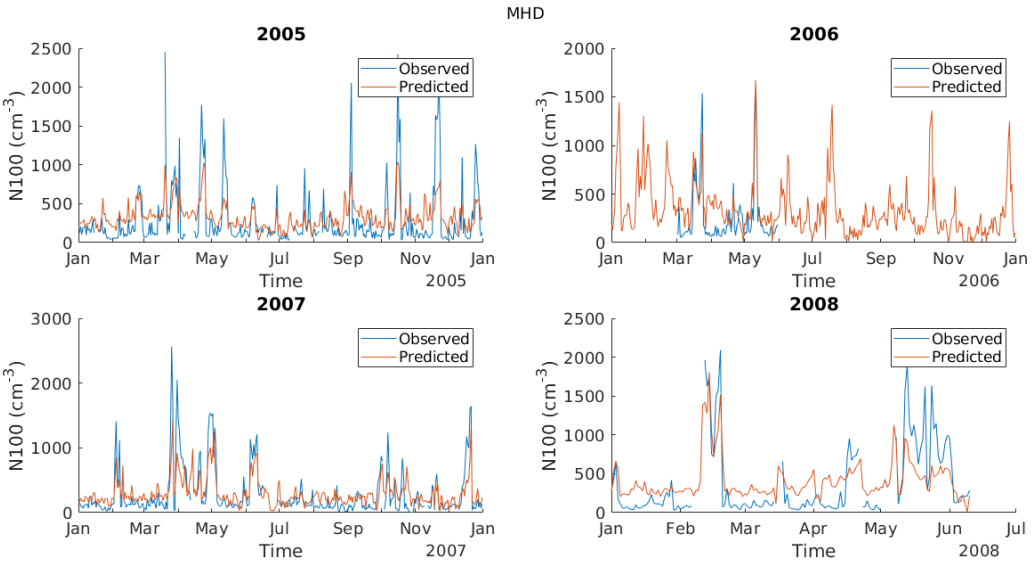

Figure 5.3: Time series of observed and predicted N_{100} for Mace Head, Ireland (MHD). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

The variation between different sites is caused by the characteristics of the locations. For example, MHD has the best R^2 value. This is probably caused by the fact that MHD is a remote location, where the normal N_{100} levels are low, but occasionally pollution is transported to the site [Nieminen et al., 2018]. During these periods, both CO and N_{100} increase, so the proxy can predict peaks quite well (figure 5.3). However, also RMSLE is quite high for MHD, which is probably caused by overestimation of low values in the proxy.

The site proxy evaluation can give similar insight also to other sites, which in turn provides information about the proxy itself. Next, observations from the site proxy at different sites are shortly discussed. The corresponding figures for each site are in Appendix A.

Some results from Alert (ALE) were already presented in regards to differences between years. Additionally, an interesting observation from Alert is how CO dominates the proxy compared to temperature. This is understandable based on the site's polar location. According to [Nieminen et al., 2018], during winter-spring, the particles are associated with long-range transport from Europe and Asia whereas clean air from the Arctic area characterises the summer air masses. As a result, the effect of biogenic aerosol is probably limited, and anthropogenic emissions explain the N_{100} concentration increase

in winter and spring. However, even though CO concentrations should represent the anthropogenic emissions quite well, the proxy does not replicate observations completely. Especially during spring 2013 the predicted N_{100} follows the shape of CO, but at the same time observed N_{100} has a clear peak (fig. 5.2) that does not correlate with CO or temperature so proxy is not able to capture it.

In Annaberg-Buchholz (ABZ), the proxy seems to work quite well, though it overestimates the lowest values, especially during winter when the concentrations are higher.

Aspverten (ASP) also has trouble finding the lowest values, but more importantly, the proxy systematically misses the high peaks throughout the year. At least some of these peaks seem to correspond to higher SO_2 concentrations, which do not directly correlate with CO. [Tunved and Ström, 2019] recognised SO_2 as an important aerosol precursor in Aspverten and the decreased SO_2 concentrations have also reduced the aerosol number concentrations during past decades. Another observation from Aspverten is that in late summer and early autumn the proxy tends to overestimate values, and this effect is more prominent during later years.

Botsalano (BOT) has issues with data availability, where both years shown in figure 5.1 have a small amount of data in the test sets, which makes the evaluation less reliable. One year is omitted as a test set because the corresponding train dataset is small. When fitting to N_{100} and CO data, the p-values for all the bins are too high and the proxy cannot be calculated. Therefore the R^2 and RMSLE values for BOT estimate proxy performance during spring and summer instead of the entire year.

In Bösel (BSL), the proxy performs well. For the most part, it can replicate observed N_{100}, though it mildly underestimates higher peaks and overestimates the lowest concentrations, especially during the winter-spring period. The site has a strong contribution from gaseous ammonia and organic emissions to SOA [Birmili et al., 2016].

The dataset used in this thesis contained only 356 days of data for Egbert (EGB) during two years. The first year covers the time period May-December, whereas the second year includes only January-April. When the proxy is trained mainly with spring data and compared to summer and autumn observations, the proxy heavily underestimates the concentrations. This shows how important it is to train and test the proxy with data covering the entire seasonal cycle. Despite the different seasons of the two iterations, in this case, both iterations of the proxy happen to yield quite similar R^2 and RMSLE.

In Helsinki (HEL), the internal variation is quite large compared to other sites. As previously discussed, this is mostly related to differences between years. In this case, there seems to be a correlation between increasing summer median temperature and the proxy performance when that year is test set (figure A.8). Overall, the proxy tends to overestimate low values, especially during winter-spring. In summer, proxy seems to exaggerate the effect of CO on the proxy. When CO concentrations are elevated, the proxy

predicts peaks with increased frequency and overestimates their height. During autumn, the proxy finds the N_{100} baseline concentration well but then misses peaks easily.

Hohenpeissenberg (HPB) is 300 m above the surrounding area which is covered with agricultural pasture and forests [Birmili et al., 2016]. It is one of the sites where the proxy struggles to capture observed N_{100}. Proxy both underestimates higher N_{100} concentrations and overestimates low concentrations. Part of this may be related to boundary layer height. For example, especially low concentrations can be observed when the site is outside the boundary layer and the instrument measures cleaner air masses of free troposphere. At the same time CAMS reanalysis CO and temperature are calculated for 10 m above ground, which is inside the boundary layer, resulting in completely different prediction. Low boundary layer heights are more common during winter, and it seems that because of this the proxy performs better during summer. Also biogenic emissions are more prevalent in summer time, which might contribute to different performances between seasons.

In Hyytiälä (HYY), the proxy captures the N_{100} relatively well, though it again has trouble with peaks and lowest values. The low concentrations are overestimated especially during summer.

In Kosetice (KCE), the proxy tends to overestimate N_{100} concentrations and cannot replicate lower values. This may be related to how proxy estimates the effect of anthropogenic emissions, which does not allow the concentrations to go low enough, even with relative error method. Kosetice is surrounded mainly by agricultural land and forests [Zíková and Zdímal, 2013]. The overestimation is more prevalent during later years.

K-Puszta (KPZ) is quite similar to Kosetice. The proxy overestimates N_{100} concentrations, especially during winter and spring. During summer, the proxy estimates roughly correct mean concentration but cannot capture the variation.

Mace Head (MHD) was shortly discussed before. The proxy mostly follows CO concentrations and the contribution of temperature is small. Mace Head is additionally one of the sites where the relative error method works less well than least squares fit when calculating the relation between CO and N_{100}. The relative error method underestimates N_{100} at high CO concentrations, so using least squares fit might capture the height of the N_{100} peaks better.

Marikana (MAR) performs relatively well during the winter months (June-August) when the N_{100} concentrations are highest. In Marikana the emission sources are metal refineries, which produce SO_2, and domestic heating and cooking [Nieminen et al., 2018]. During summer, however, the proxy is only able to capture average concentration, but not the variation, and peaks are underestimated.

In Melpitz (MLP), the proxy captures observed N_{100} well, especially in summer months. During winter-spring, the proxy overestimates lower values which decreases the

overall performance.

Nanjing (NAN) is similar to Marikana in that it is an urban site, where the predicted N_{100} follows observations better during winter. In summer, the proxy is able to estimate the average concentrations but struggles to capture the variation. Additionally, it occasionally overestimates the peaks.

In Neuglobsow (NEU), the proxy captures well the observed variation in N_{100}, though it does not quite manage to cover the highest and lowest values.

The dataset used in this thesis contains only 283 days of data for Saõ Paulo (SAO). The days are split between two years so that 2010 has 71 days of data between October-December and 2011 has 212 days of data from January to September. Training the proxy with 2010 data leads to parameters and evaluation results that are clearly different compared to a larger training set. Therefore, only results using 2011 as a training set are shown in figure 5.1. However, because of this, the test set is only 71 days making the analysis less reliable. Furthermore, the proxy is trained with data from summer to winter and then tested against observations from spring-summer. Based on other urban sites like Nanjing and Marikana, the site proxy has difficulty in replicating observed N_{100} during summer. Therefore, the real proxy performance is probably better than this result indicates.

Schauinsland (SCH) is the second mountainous site, similar to HPB. While the proxy performs better in SCH than in HPB, for the most part it can capture only the average concentration and misses a significant portion of the variation. The proxy performance is better during summer and autumn, but the correlation between observed and predicted N_{100} is poor during winter and spring. This is probably related to the boundary layer height, since according to [Birmili et al., 2016] Schauinsland is typically above the inversion layer in winter.

Southern Great Plains (SGP) is another site where the proxy performs less well since the proxy can not replicate the highest and lowest values.

In Vielsalm (VIE), the proxy performs relatively well, but it struggles with capturing low concentrations.

In Värriö (VAR), the proxy performance varies between years. The concentrations are higher during summer, and while the proxy does underestimate the highest concentrations, the correlation between observed and predicted N_{100} is good. During winter and spring, however, the proxy estimates average concentration but does not capture the variation. Additionally, unlike in other sites, in Värriö the proxy occasionally significantly underestimates the concentrations, especially in summer and autumn.

Waldhof (WAL) is similar to many other European sites. The proxy works especially well during summer and autumn, but it has a tendency to overestimate low concentrations all around the year.

In summary, for the site proxy, the R^2 values for continental sites vary between 0.12 and 0.68 and the RMSLE values between 0.16-0.31. While the proxy performs differently depending on location, in general the lower values are overestimated. The variation between years is also significant.

5.1.2 The site excluded proxy

Figure 5.4 shows how proxy performs when the site's own data is excluded from training and only other sites' data is used. Ranges were attained from using different train and test sets to evaluate how much data selection affects the results. Same as in the site proxy, the train-test selections, where test sets were small and results clearly deviated from other selections, were omitted from the figure. Omitted test sets were otherwise the same as in figure 5.1, except for SAO, where 2011 data was used as a test set instead of 2010.

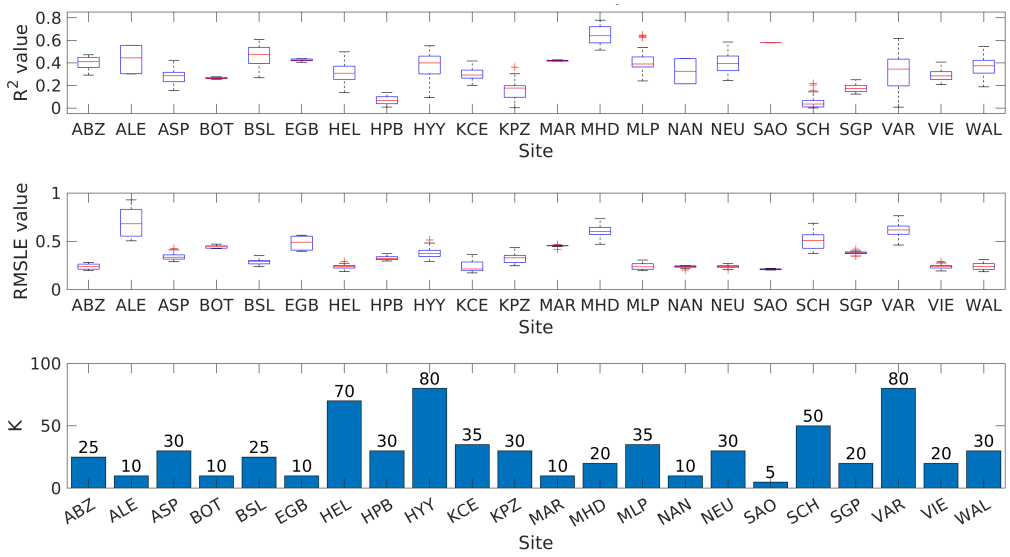

Figure 5.4: Figure shows how well the site excluded proxy is able to predict N_{100} when trained with other sites' data. The train and test sets were varied by selecting always one year of site's data as test set and randomly sampling five train sets from other sites' data. Boxplots estimate the influence different data selection has on proxy performance. Red line shows median R^2 and RMSLE values, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and red crosses outliers. Additionally K value describes how many data points is included in boxplots.

Comparing the site excluded proxy (fig. 5.4) to the site proxy (fig. 5.1), R^2 values are lower and RMSLE higher suggesting overall poorer performance. Especially HBP, KPZ, and SCH suffer, indicating that parameters calculated from other locations' data cannot capture the processes properly. For HBP and SCH this might relate to high elevation and mountainous characteristics. Furthermore, even though R^2 remains quite good for ALE, the proxy overestimates N_{100} causing large RMSLE. This is probably associated with ALE being a polar remote location, where the biogenic emissions are limited and

anthropogenic emissions are from long-range transport, causing different CO and N_{100} dynamic compared to other locations closer to sources. Similar observations can also be made in other clean or remote sites, like MHD, SCH, and VAR.

The sites, where the proxy improves, are HEL and SAO. For HEL this is probably just a coincidence, but for SAO it is caused by different test set selection. Previously in the site proxy, the 2011 test set was excluded because the corresponding train set was too small to calculate the parameters. Here this is no longer an issue as the proxy is trained with data from other sites. Additionally, the previously used 2010 test set contained only 71 data points, which made it less reliable, so here 2011 is selected as a test set instead. With this test set the result improves significantly compared to the site proxy.

Another difference between the proxies is that using data from other sites emphasises the effect of train and test set split, causing larger variation within sites (fig. 5.4). For some sites, this is more related to test sets, and regardless of the train set the proxy results are better or worse for some years than others. These include ALE, where the proxy trained with other sites' data is better at predicting N_{100} for 2013. Also, NAN and EGB have clear differences between years. For NAN, the reason behind this is not known, but for EGB, it is probably caused by limited data coverage. EGB has data from the period of two years. The first year, 2007, starts from May, covering mostly summer, autumn, and early winter. Data from 2008, on the other hand, covers the period from the start of the year to May. While this does not affect R^2 values much, RMSLE is higher for the spring months.

For other sites, the increased variation is more related to train data selection. This is true even when considering only one test set and comparing results from different random train sets. However, making a direct comparison between the site proxy and the site excluded proxy is difficult since the number of data points in each boxplot is five times larger in the site excluded proxy.

In summary, the continental sites have R^2 values ranging from 0.0 to 0.61, and RMSLE values from 0.17 to 0.92. Compared to the site proxy, the performance is decreased, with most sites having lower R^2 values, larger RMSLE values, and larger variation between years. Of the continental sites, especially in KPZ the performance decreases compared to the site proxy.

5.1.3 Global proxy

Finally, the evaluation results for the global proxy can be seen in figure 5.5, including the ranges from different train-test set splits. Small test sets with outlier results were again manually excluded from the figure. Overall, the results are similar to the site excluded proxy (fig. 5.4) discussed in the previous section, though adding the site's own data

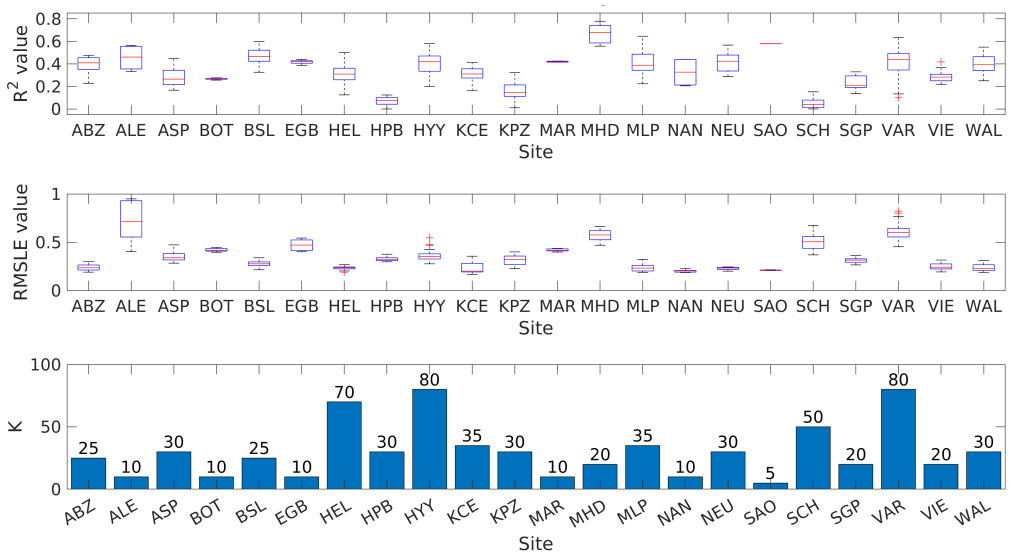

Figure 5.5: Results from the global proxy, where N_{100} is predicted with proxy trained with one year of data from all continental sites. The train and test sets were varied by selecting always one year of site's data as test set and randomly sampling five train sets from all continental sites. Boxplots estimate the influence different data selection has on proxy performance. Red line shows median R^2 and RMSLE values, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and red crosses outliers. Additionally K value describes how many data points is included in boxplots.

to the training set both narrows the variation within each site and slightly betters the results. Compared to the site proxy (fig. 5.1), global proxies perform moderately worse in most locations, though the differences are typically small. The exception to this are again HBP, KPZ, and SCH, where the global proxy performance is significantly lower, and SAO, where the improvement is related to test set change.

In summary, for continental sites the evaluation of the global proxy gives R^2 values from 0.01 to 0.64 and RMSLE values between 0.17 and 0.95. The results are very similar to the site excluded proxy.

5.2 The global proxy

The final global proxy was generated by training 100 proxies with different global datasets and then producing and evaluating the predicted N_{100} using the entire available data for the site. The results for 100 generated global proxies can be seen in figure 5.6 and the distributions of generated parameters in figure 5.7.

As expected, the final global proxy performance is quite similar to the global proxy evaluations from the previous section. The main exception here is that since the final global proxy is tested against all available data from each site instead of individual years, the effect of yearly variations is reduced. Therefore, the final global proxy tends to have smaller variations at each site. However, it should be noted that in Alert (ALE), some of

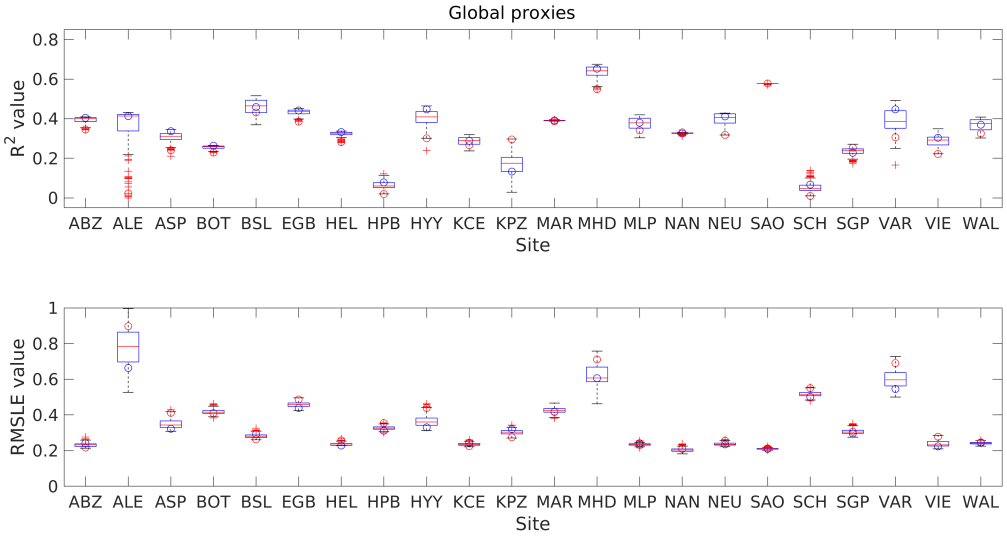

Figure 5.6: R^2 and root mean squared log error (RMSLE) from the 100 global proxies trained with randomly generated datasets containing one year of data from each continental site. Red line shows median, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and red crosses outliers. All boxplots contain 100 datapoints. Additionally, blue and red circles show the two global proxies selected as examples. The mode parameter proxy is indicated with blue and the comparison proxy with red.

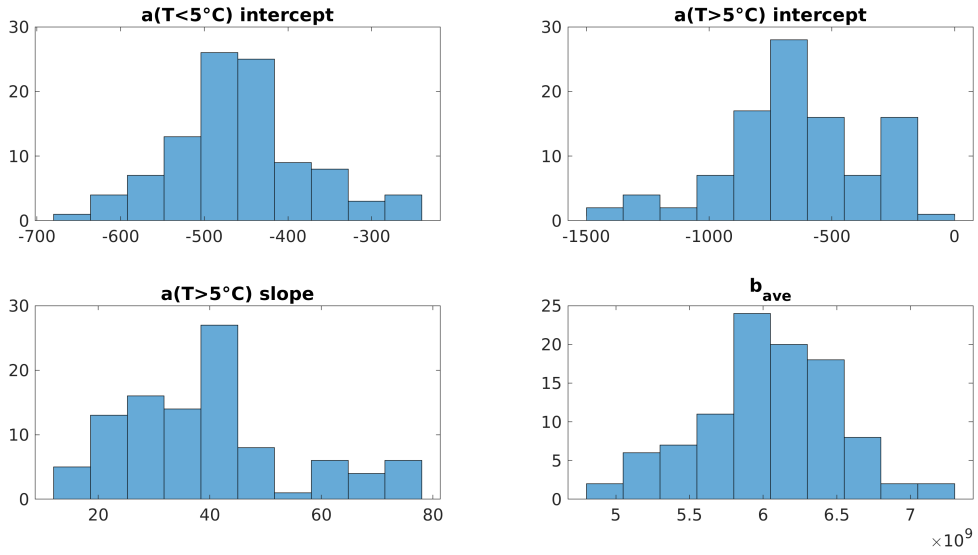

Figure 5.7: Distributions of parameters from the 100 global proxies.

the global proxy versions perform less well than would have been expected based on the global proxy evaluation.

Aside mountainous sites HPB and SCH performing poorly, it is not directly clear that any type of site would perform better or worse compared to others. The only other conclusion that can be made based on site types is that in cleaner and more remote

sites RMSLE values tend to be higher. For further analysis, proxy performance was also compared against selected variables from CAMS, including temperature and CO, but also other trace gases like SO_2, NO_x and terpenes. This was done by selecting 10 best R^2 and RMSLE values for each site and looking at the correlations between medians of the evaluation metrics and the medians of the variables for each site. Based on this analysis, there is no clear connection between site performance and the meteorological or gas variables. R^2 values do not correlate with the variables. For RMSLE it is possible that certain conditions decrease the error, namely RMSLE seems to be lower for warmer sites (R = -0.59) and also if NO_x/CO relation is high (R = -0.63) (figure not shown).

Next, the proxy parameters and their effect on performance was considered. A more detailed inspection of the global proxy parameters (fig. 5.7) shows that the parameters can vary quite radically depending on what dataset is used for training. Regardless, all parameters have a distinct mode value. A global proxy that has all parameters inside the maximum bin was selected for further analysis. Figure 5.6 shows the results for this "mode parameter proxy" with blue circles and the parameters are shown in table 5.1.

Table 5.1: Global proxy parameters for the two selected global proxies.

Name	a(T<5°C) intercept	a(T>5°C) intercept	a(T>5°C) slope	b_{ave}
Mode parameter proxy	-487.02	-642.70	39.07	5.930e+09
Comparison proxy	-466.41	-269.19	16.61	5.932e+09

While figure 5.7 shows the frequency of different parameter values in the global proxy when the train set is varied, it does not tell anything about how the sites are affected. Therefore, figure 5.8a shows the range of the ten parameters that yield the highest R^2 for each site and, similarly, figure 5.8b has the ten parameters that produce the lowest RMSLE. Additionally, the blue line indicates parameters for the mode parameter proxy. Predictably, the best parameters for each site are different and also vary depending on which metric is used. There are some similarities in the parameters for same type of sites, but for the most part, this is not clear enough to make definitive conclusions.

For further analysis, dendrograms that cluster sites together based on the parameters that produce the best R^2 value (fig. 5.9a) and the best RMSLE value (fig. 5.9b) were created. Clustering was done using Matlab *linkage* function, where the clusters were calculated based on the euclidian weighted center of mass distance. This method can find some groups within the sites. For example, in figure 5.9a red cluster contains Central-European rural sites, the purple cluster also has rural sites KCE, KPZ, and SGP. However, the rest of the clusters are more difficult to explain from a physical standpoint. Namely, HPB and SCH have understandably similar parameters, as they are mountainous sites, but having them in the same branch as MAR and VIE is physically less reasonable. Similarly,

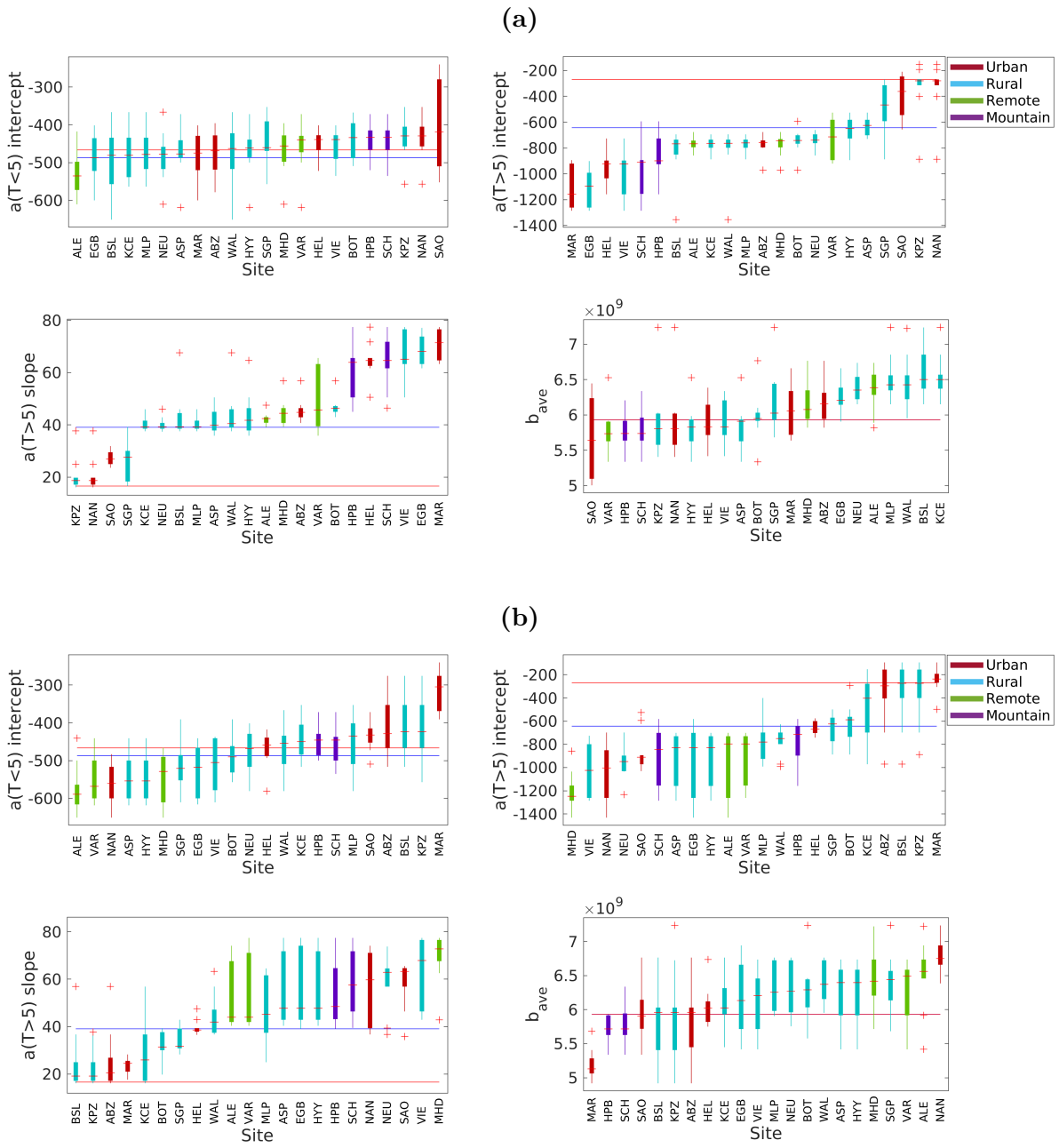

Figure 5.8: Panel a) shows the parameters for the 10 global proxies that produce largest R^2 values for each site, and b) same for smallest root mean squared log errors. Boxplot colors indicate the site type. Blue and red lines show the parameters for the two selected global proxies.

for dendrogram based on RMSLE (fig. 5.9b) ALE and VAR have similar parameters due to their arctic and remote locations, and HPB and SCH have the same parameters again, but for instance, the blue branch contains a variety of different types of sites. All in all, it seems that the parameter ranges that produce well-performing proxies are large, and grouping sites together based on their parameters is challenging. Parameters were also compared against selected CAMS variables and their 25th-75th percentiles, but no

connection between variables and parameters was found.

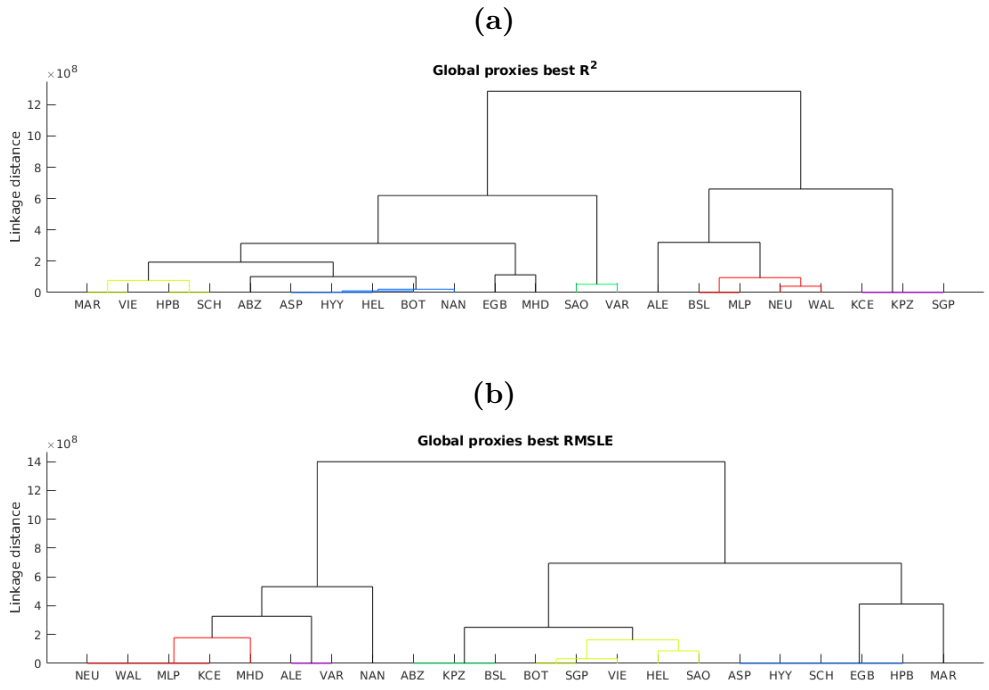

Figure 5.9: Figure shows the dendrograms produced with Matlab linkage function using euclidian weighted center of mass distance as clustering method. Panel a) shows sites clustered together based on parameters that produces largest R^2 values for each site, and b) same for smallest root mean squared log errors.

On the other hand, if the parameters are not distinguishable by site type, it might be possible to find a parameter set or sets that perform reasonably well globally and not only at a certain type of location. For example, the fact that in figure 5.9b the linkage distance between sites in the blue cluster is zero indicates that for all those sites the same parameter set produces the smallest RMSLE. While no one parameter set would produce the best results for all sites, e.g. the mode global proxy parameters are within most of the sites' best parameter ranges (fig. 5.8). To further examine the effect parameters have on the global proxy, a version of global proxy with significantly different parameters was selected. This "comparison proxy" is indicated with red line in figure 5.8 and red circles in figure 5.6. The parameters can be seen in table 5.1. The comparison proxy differs from the mode parameter proxy mainly due to a(T)-parameter. The comparison proxy has a smaller slope at temperatures above 5°C and, hence also a larger intercept. Based on this, the comparison proxy should predict a smaller increase in N_{100} with increasing temperature compared to the mode parameter proxy. Next, these two parameter sets and their effect on predicted N_{100} are compared.

As figure 5.6 shows, depending on site the difference between the two selected prox-

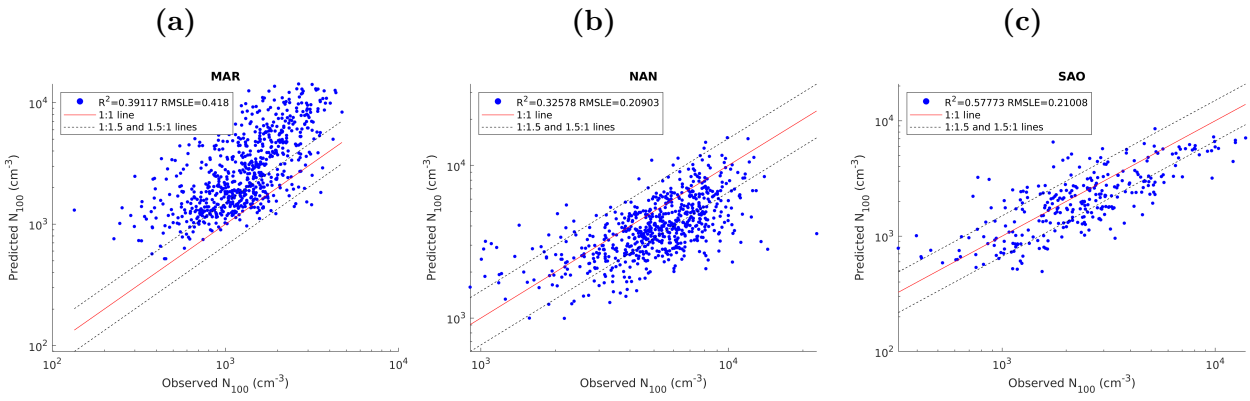

Figure 5.10: Predicted N_{100} based on the mode parameter global proxy compared to observed N_{100} for a) Marikana b) Nanjing and c) Sao Paulo. These are sites where the proxy results are not affected by parameter selection.

ies with different parameters can be insignificant or very large. There seems to be sites like MAR, NAN, and SAO where the parameters do not affect the result. This is true for the two selected parameter sets but also all the 100 global proxies. The commonality between these sites is that they have a relatively short dataset, which might contribute to small variation. However, they are also urban sites with large anthropogenic emissions. It is possible that in these locations the global proxy follows the CO-dependant part and, since b_{ave} varies proportionally less than other parameters, this produces the same results regardless of parameters. Looking at how the proxy performs in these sites, the comparisons between observed and predicted N_{100} from the global proxy with mode parameters can be seen in figure 5.10. In MAR (fig. 5.10a), the predicted N_{100} is heavily overestimated with days up to ten times too high concentrations. On the other hand, for NAN (fig. 5.10b) the global proxy tends to underestimate N_{100}, and from time series it can be seen that underestimation mainly occurs during spring and summer (figure not shown). In SAO (fig. 5.10c), the proxy performs well. All available data from SAO is used as a test set, so the result should show the global proxy performance covering almost the entire seasonal cycle. However, since SAO has only this one year of data, it is also used in training the proxy. Since training and testing are partially done with the same data, there is a small risk that this boosts proxy performance. Although, since the site's own data is only around 1/19 of the training set with all continental sites, this should not be an issue.

For most of the other sites, the proxy performs better with mode parameters than the comparison parameter set. The only exception to this is KPZ. As mentioned before, typically KPZ does not respond well to parameters trained with other sites' data, but with the comparison parameters, the proxy performance is significantly better. While

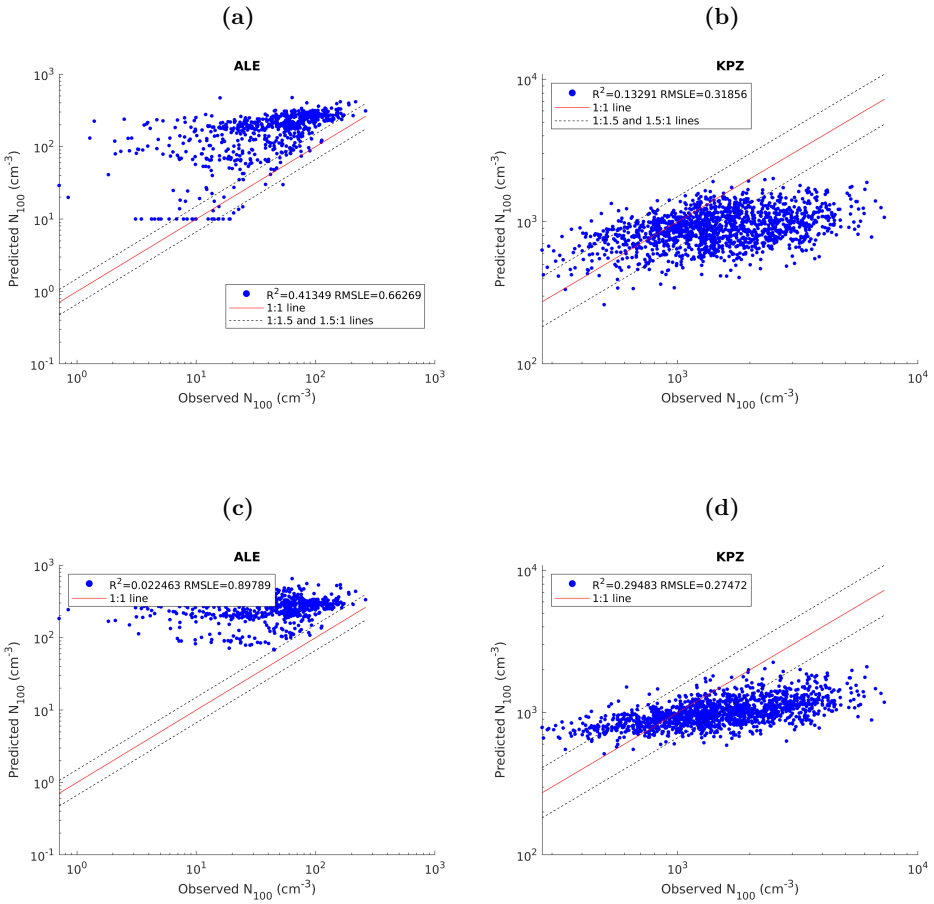

Figure 5.11: Predicted N_{100} from the global proxy compared to observed N_{100} for Alert (ALE) and K-Puszta (KPZ), where in panels a)-b) the predicted N_{100} are calculated with the mode parameters and in panels c)-d) with the comparison parameters (table 5.1).

the difference in performance on most of the other sites is not large, an exception to this is ALE where using the comparison parameters significantly reduces the proxy performance. Figure 5.11 shows the comparisons between observed and predicted N_{100} for ALE and KPZ with both parameter sets. Looking first at ALE, even with the mode parameters (fig. 5.11a), the proxy systematically overestimates N_{100} concentrations, and especially low concentrations are predicted poorly. There is also a group of data points with 10 cm^{-3} concentration which is a result of setting a lower limit for the proxy to prevent negative values. For the comparison parameter set, the proxy overestimates N_{100} concentrations completely and the correlation between predicted and observed N_{100} is very poor (fig. 5.11c). Next, regarding KPZ, it can be noted that while the correlation is better with the comparison parameter set, both proxies underestimate the higher concentrations (figs. 5.11b and 5.11d). Therefore the proxy is not able to, for example, capture the seasonal cycle of observed N_{100}. This demonstrates the difficulty in producing a global proxy. Firstly, it is challenging to find parameters even for one site that could replicate

observations. Secondly, the parameters that produce the best results for one site cannot necessarily do this for other sites, as ALE and KPZ clearly illustrate.

Next, the rest of the sites are shortly discussed and the comparisons between predicted and observed N_{100} are shown, concentrating mostly on the mode parameter global proxy, as it produced better results. Sites can be roughly divided into four groups: sites where the global proxy predominantly underestimates concentrations, sites that are overestimated, a combination of these, and mountainous sites.

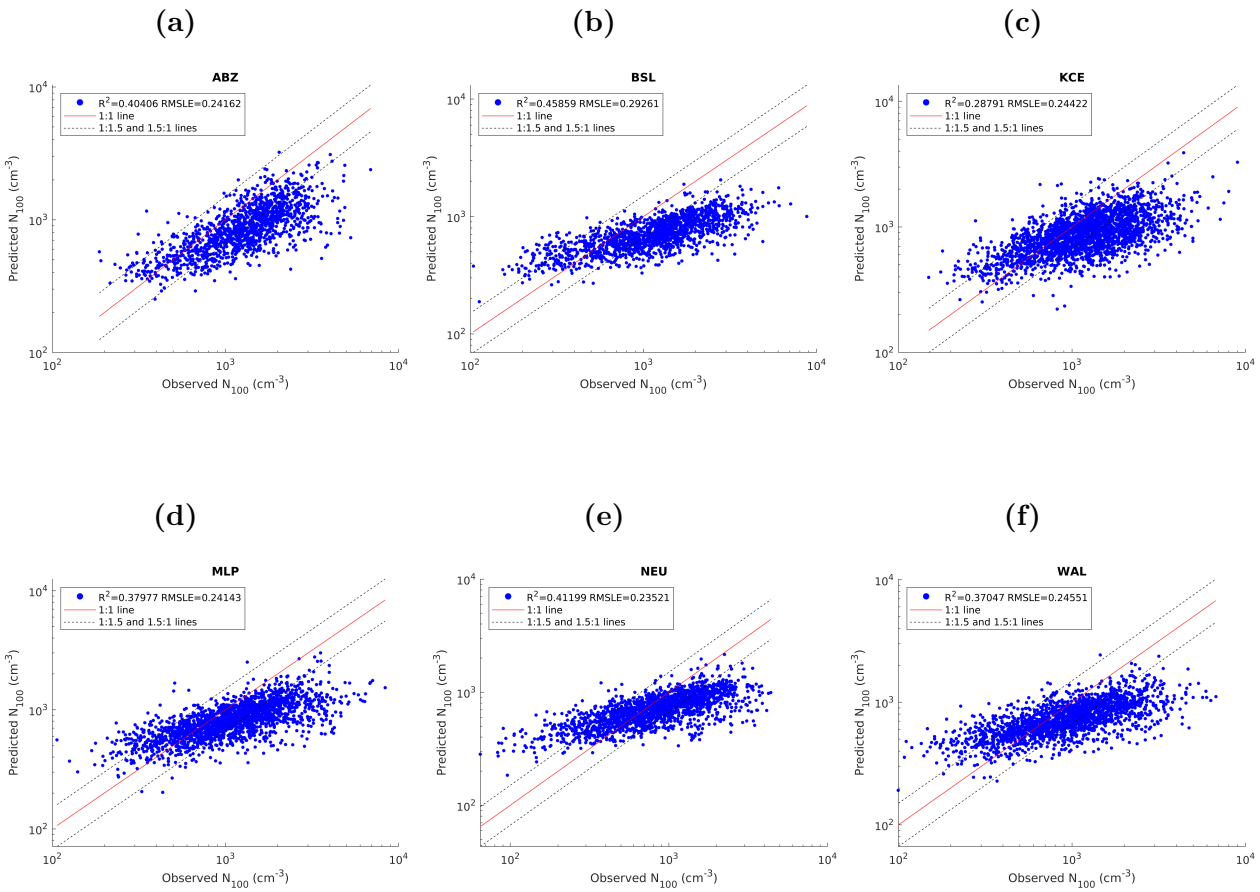

Figure 5.12: Predicted N_{100} from the mode parameter global proxy compared to observed N_{100} for a) Annaberg-Buchholz (ABZ), b) Bösel (BSL), c) Kosetice (KCE), d) Melpitz (MLP), e) Neuglobsow (NEU), f) Waldhof (WAL). These are sites where proxy underestimates higher N_{100} concentrations.

First, the sites where the proxy underestimates N_{100} when the concentrations are high include ABZ, BSL, KCE, MPL NEU, and WAL. In ABZ, this underestimation is relatively small and for the most part, proxy performs well (figure 5.12a). In BSL, MLP NEU and WAL, however, the proxy is able to capture the seasonal cycle but underestimates heavily higher concentrations (figures 5.12b and 5.12d-5.12f). KCE (figure 5.12c), on the other hand, has higher N_{100} concentrations during the first couple of years, and the proxy underestimates these, but when the observed concentrations decrease proxy starts

to perform moderately better. Excluding the first years increases R^2 to 0.30 and lowers the RMSLE to 0.19, though the proxy still underestimates slightly. Interestingly, all of these sites are located in Central Europe, and with exception of ABZ, are rural sites. While ABZ is an urban site, it is an urban background site.

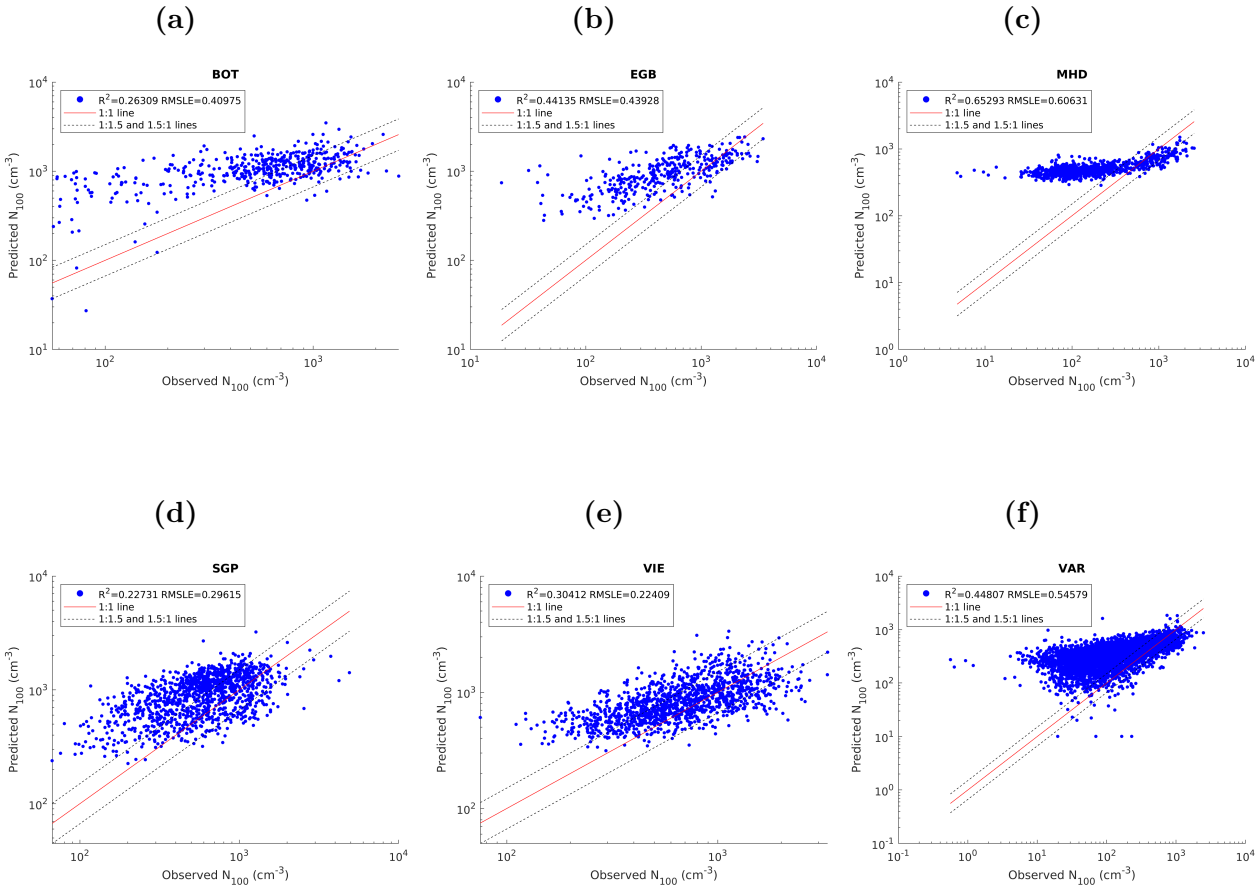

Figure 5.13: Predicted N_{100} from the mode parameter global proxy compared to observed N_{100} for a) Botsalano (BOT), b) Egbert (EGB), c) Mace Head (MHD), d) Southern Great Planes (SGP), e) Vielsalm (VIE), f) Värriö (VAR). These are sites where proxy overestimates lower N_{100} concentrations.

The second group contains BOT, EGB, MHD, SGB, VIE, and VAR, where the low concentrations are overestimated (figure 5.13). Also, ALE, which was already discussed, belongs to this group (figure 5.11a). In Botsalano (fig. 5.13a) the overestimation occurs especially during summer when the proxy is not able to replicate the variation in N_{100} concentrations. For Mace Head, the global proxy is still good at finding peaks, but it strongly overestimates low concentrations, producing a clear tail to figure 5.13c and also results in high RMSLE (figure 5.6). In SGP, in addition to overestimation, based on time series it seems that the proxy follows too strongly temperature, resulting in a more emphasised seasonal cycle compared to observations. The overall commonality in these sites is that they have N_{100} concentrations that reach values below 500 cm^{-3}. It seems

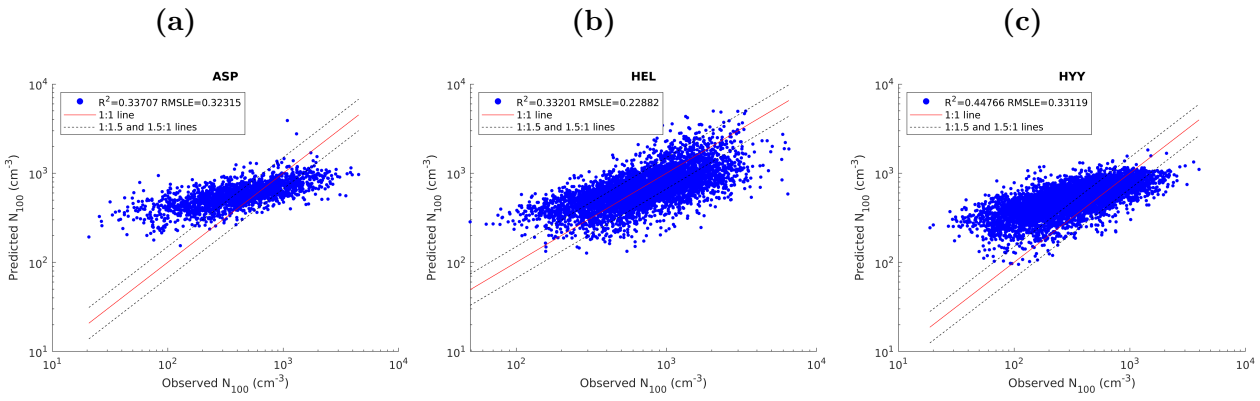

Figure 5.14: Predicted N_{100} from the mode parameter global proxy compared to observed N_{100} for a) Aspvreten (ASP), b) Helsinki (HEL), c) Hyytiälä (HYY). These are sites where proxy both underestimates higher N_{100} concentrations and overestimates lower N_{100} concentrations.

that while the proxy can replicate concentrations lower than 500 cm^{-3}, it struggles to do so reliably.

Third, there are the sites that both overestimate low concentrations and underestimate high concentrations, such as ASP, HEL, and HYY (figure 5.14). Of these, ASP and HYY are rural sites in Sweden and Finland, whereas HEL is an urban site in Finland. In all these sites the proxy can capture the seasonal variation but fails to replicate low and high concentrations. For example, in HEL during summer the proxy replicates high peaks quite well, but during winter similar peaks are underestimated. As before, the overestimation of low values is related to challenges at capturing concentrations below 500 cm^{-3}.

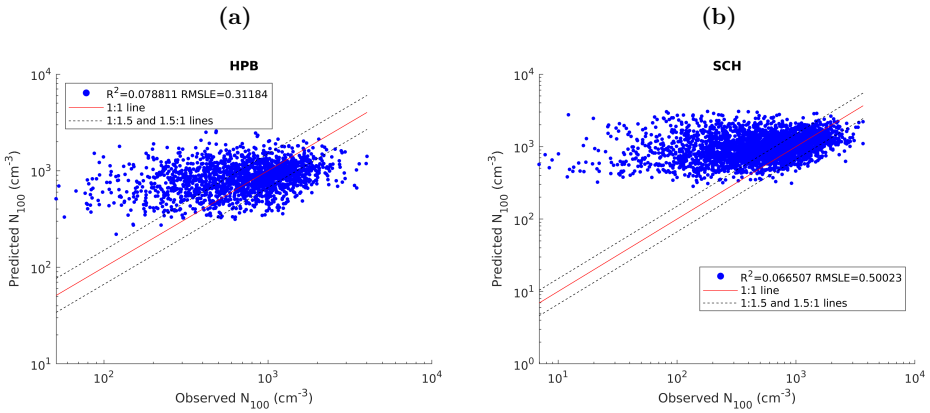

Figure 5.15: Predicted N_{100} from the mode parameter global proxy compared to observed N_{100} for a) Hohenpeissenberg (HPB), b) Schauinsland (SCH). These are mountainous sites with poor proxy performance.

The final group contains mountainous sites HPB and SCH (figure 5.15). In Hohen-

peissenberg the correlation between observed and predicted N_{100} is very poor. Further analysis of the time series reveals that especially during winter the predicted N_{100} is quite high, likely due to anthropogenic emissions from e.g. heating, and at the same time observed N_{100} is very low probably because it is outside boundary layer. On the other hand, in summer the observed concentrations are higher, but proxy predicts relatively low values. All in all, it is clear that the proxy does not work in Hohenpeissenberg. Part of this might be related to mountainous sites, including HPB, being excluded from training data. However, since the site proxy does not perform well in HPB either, it is possible that temperature and CO are not the best variables to describe N_{100} in a site like this. Similar observations hold also for SCH, where it is known that the site is outside the boundary layer during winter.

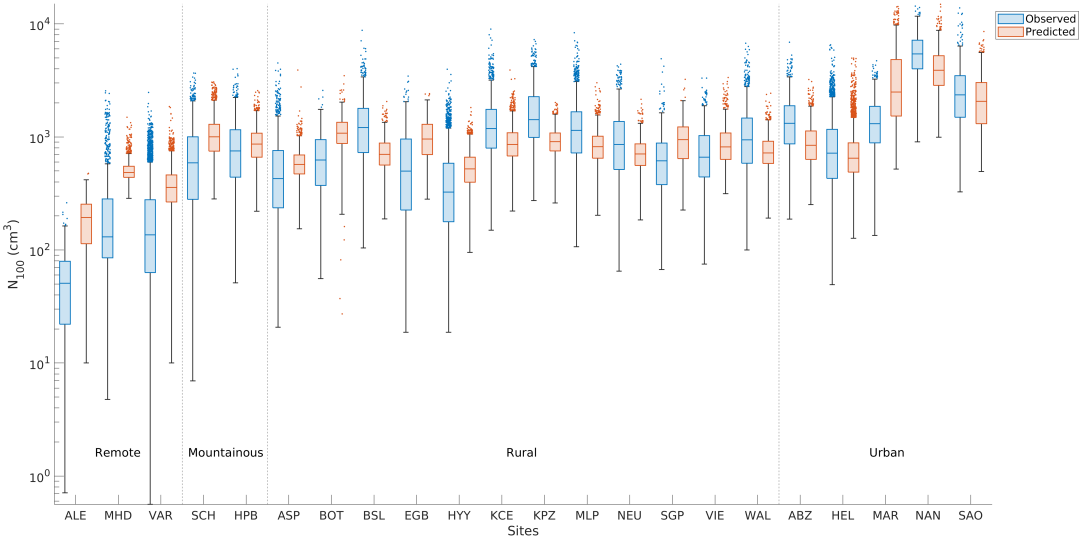

Figure 5.16: Observed N_{100} ranges and predicted N_{100} ranges from the mode parameter global proxy. Line shows median N_{100}, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and dots outliers.

Finally, figure 5.16 shows the total observed N_{100} range at each site and the corresponding predicted values from the mode parameter global proxy. If observed data has gaps, also predicted N_{100} has been removed to ensure a fair comparison. Additionally, in figure 5.17 the data is further divided into seasons. Figures illustrate well the large differences in concentrations between sites from cleaner remote locations to polluted urban sites. Moreover, as already noted previously, at many locations the proxy generally overestimates N_{100} concentrations compared to observations. This occurs especially in remote clean locations like ALE, MHD, and VAR, as well as South-African sites MAR and BOT, but also in SCH, EGB, and SGP. In MAR this is emphasised during December-February 5.17a. Significant underestimation occurs mostly in European rural sites like BSL, KCE,

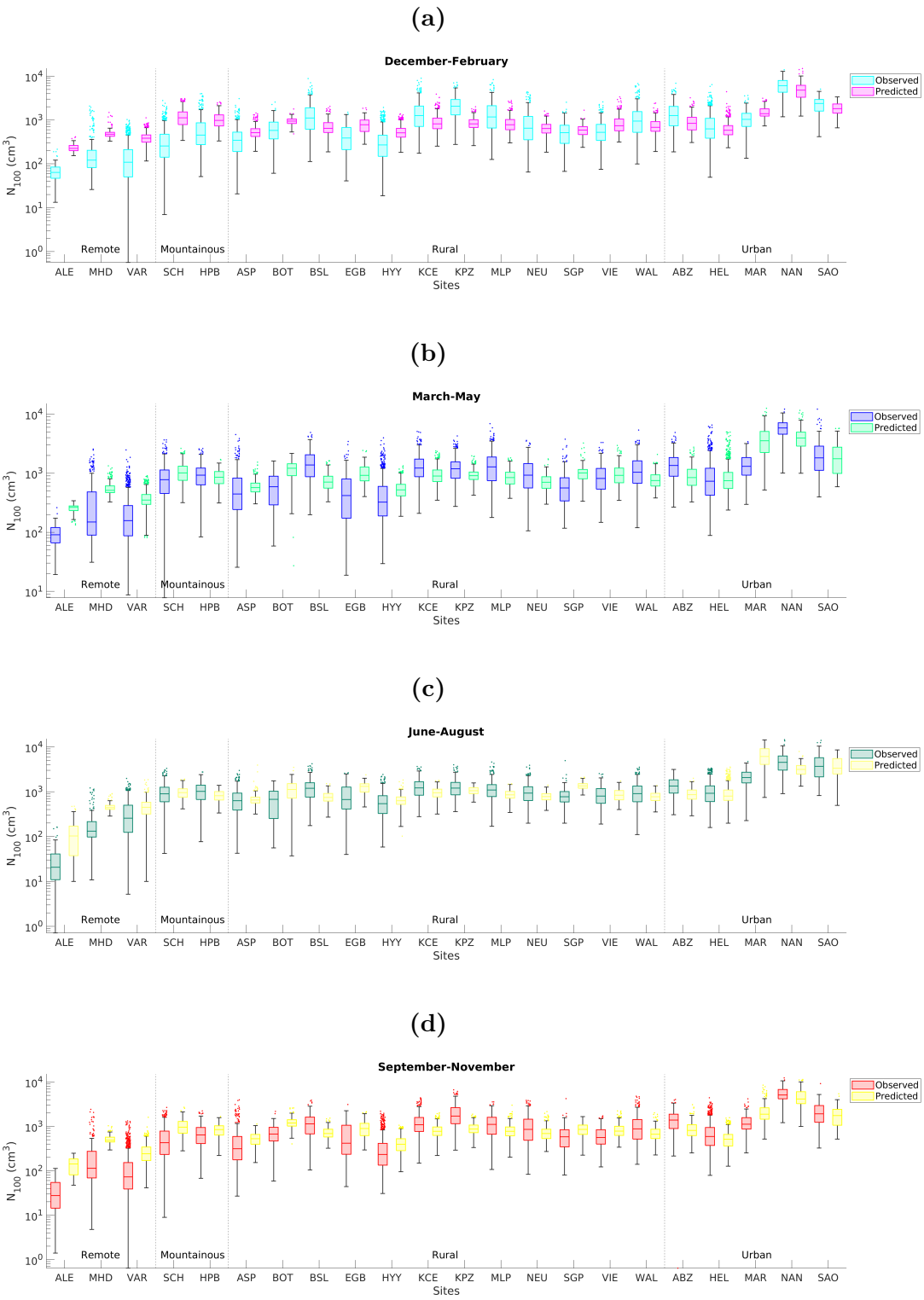

Figure 5.17: Observed N_{100} ranges and predicted N_{100} ranges from the mode parameter global proxy divided by season. Line shows median N_{100}, box 25th and 75th percentiles, whiskers datapoints within 2.7σ range and dots outliers.

KPZ, and additionally in NAN. In NAN the underestimation takes place mostly during March-May. For the rest of the sites, the proxy does not capture the entire variation but the predicted values are in the correct range. The median concentrations are predicted most accurately at VIE and SAO. Typically the proxy is better at capturing the variation in N_{100} during spring and summer, but this depends on site.

In figure 5.18 the oberved and predicted median N_{100} concentrations are further compared. It can be seen that sites, where the median N_{100} concentration is below approximately 500 cm^{-3}, are typically overestimated. Sites with higher median concentrations are closer to 1:1 line. In SAO and NAN, which are the most polluted urban sites in the dataset, the proxy replicates median N_{100} well, but some of the sites with median concentrations around 1000 cm^{-3} are slightly underestimated. The exception to this is MAR, which is heavily overestimated. When looking at the seasonality, the proxy seems to capture the median N_{100} better during June-August, which is the northern hemisphere's summer. Of the sites in southern hemisphere, BOT and SAO are not much affected by the seasonality, but MAR also performs better during summer months.

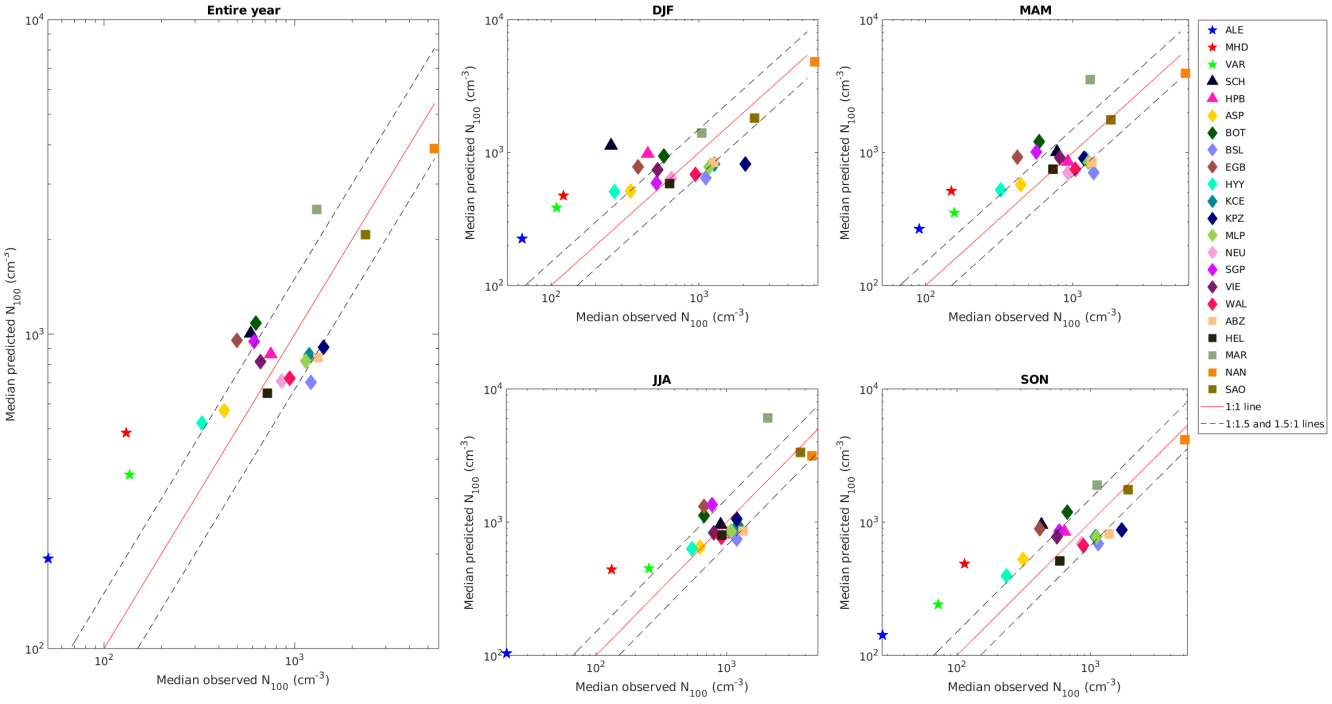

Figure 5.18: Observed median N_{100} concentrations compared against predicted median N_{100} concentrations for each site. Left panel shows results for entire year and right panels for each season. Remote locations are indicated with star, mountainous sites with triangle, rural with diamod and urban with squares.

6. Discussion

The main aim of this thesis was to produce a global proxy that predicts N_{100} based on temperature and CO from CAMS data. While the work done in this thesis shows clearly that temperature and CO can be used as tracers for N_{100}, the result for the global proxy has several limitations. Firstly, [Rosenfeld et al., 2014] states that for assessing the effect aerosols have on clouds and precipitation, the CCN needs to be globally estimated with accuracy within a factor of 1.5. As the global proxy results (figures 5.10-5.15) show, the method developed in this thesis is not able to reach this accuracy. The proxy has best accuracy in SAO with most days withing the 1.5 limit, but even there the proxy tends to underestimate. Good accuracy would be required especially in clean areas [Rosenfeld et al., 2014], but currently the global proxy struggles to capture low concentrations. This results in high RMSLE in sites like ALE, VAR, and MHD.

Secondly, for most sites, the global proxy is not able to reach the performance of the site proxy. Depending on site, the site proxy yielded R^2 values between 0.13 and 0.68 when evaluated against observed N_{100}, whereas for the global proxy R^2 range was 0.03-0.58. Additionally, RSMLE error values increased for the global proxy, especially at the clean sites mentioned before. The site proxy performing better than the site excluded or the global proxy is understandable, since, for example, site's distance from anthropogenic sources, locations within or outside boundary layer, and the vegetation surrounding the site, all affect how N_{100} responds to changes in temperature and CO. Hence, the parameters for each site are quite different and it is difficult to find one set of parameters that would describe all sites well. In the global proxy, this was demonstrated by the differences between the mode parameter proxy and the comparison parameter proxy. When using comparison parameter proxy, results for KPZ significantly improved compared to mode parameter proxy, but this change also significantly worsened proxy performance in ALE. On the other hand, despite the large variation in parameters within each site, for example in the 10 best global proxies (fig. 5.8), the evaluation results stay relatively constant. This indicates at least some level of resilience to changes in parameters.

Thirdly, proxies are needed to estimate CCN concentrations at locations where there are no measurements available. The site excluded proxy allowed estimating the proxy

performance in this scenario by leaving one target site out of the training set and then testing against test site's data. Comparing the site proxy and the site excluded proxy, the variation in results between sites was larger in the site excluded proxy, indicating a significant difference in proxy performance depending on location. This was also seen in the global proxy. Unfortunately, no clear division into better and worse performing sites was found based on either site type or CAMS variables, though mountainous sites systematically performed poorly and clean sites had large RMSLE values. For the site proxy, cleaner sites and some of the European rural sites performed better, whereas in the global proxy the two sites with best performance were MHD and SAO. The reason why the proxy works better at some European rural sites compared to others was left unanswered. Since the performance depends on site, extrapolating a global proxy to locations with no reference measurements would be difficult, as the accuracy could not be assessed.

Comparing the developed proxy to other earlier CCN parametrisations gives an idea of the proxy performance in the context of wider literature. Looking first at individual parameters, no research was found that would quantify linearly the effect of air temperature on particle concentrations. Instead, the emission rates between particle number and CO concentrations have been studied, mostly in the context of biomass burning. For example, [Guyon et al., 2005] estimated relation between the number of particles produced in Amazonian deforestation fires and CO emissions. Their results gave a ratio in the range of 14-32 $cm^{-3}ppb^{-1}$, which is in agreement with other studies. In this thesis the emission ratio is expressed with parameter b_{ave}. For the global proxy, parameter b_{ave} is within 4.8-7.0 when converted into $cm^{-3}ppb^{-1}$. The mode parameter corresponds to 5.7 $cm^{-3}ppb^{-1}$. These are much lower than the estimation for forest fires in [Guyon et al., 2005], though it should be noted that Guyon included particles between from 8 nm to around 300-500 nm instead of just accumulation mode, and fossil fuel combustion may have completely different emission ratio. Additionally, the atmospheric lifetime of CO is longer than a typical atmospheric particle, which lowers the particle to CO ratio when away from sources.

Next, the results from this thesis are compared to proxies and parametrisations from previous studies that have used some of the same measurement sites. A relatively recent study [Shen et al., 2019] developed CCN parametrisations based on in-situ measurements of CCN and aerosol optical properties including back-scattering fraction and aerosol scattering coefficient. One of the sites they used was HYY, for which they found R^2=0.45 at high supersaturation and R^2=0.61 at low supersaturations when comparing observed and predicted results. Another parametrisation method resulted in R^2 between 0.5-0.84. The results from this thesis give the global proxy R^2 range between 0.23-0.46 and median of R^2=0.40 for HYY whereas the site proxy for HYY has a range of 0.20-0.58 with a median of 0.42. Based on this, the global proxy results are lower compared to [Shen et al., 2019], and upper boundary of the site proxy results overlap with lower R^2

from [Shen et al., 2019].

Another study involving optical properties is [Liu and Li, 2014], where they created a CCN parametrisation for continental rural areas using in-situ aerosol index and aerosol optical depth and CCN at 0.4% supersaturation from SGP. Their method was able to achieve R^2 above 0.94 for selected conditions where single-scattering albedos were limited to 0.85-0.95 and RH<80%. In comparison, the method developed in this thesis yielded a median R^2 of 0.28 and a maximum R^2 of 0.34 for the site proxy. The results from the global proxy gave a median R^2 of 0.24 and a maximum R^2 of 0.27. Based on this, the method in [Liu and Li, 2014] performs significantly better in SGP, where the method developed in this proxy struggled compared to other rural sites. However, it is worth noting that using optical properties for to global estimates is still challenging. [Stier, 2016] evaluated that daily CCN predictions based on aerosol index yield $R^2 = 0.46$ over continental and marine regions. If marine regions are excluded and the averaging is done over longer time scale the R^2 increases to 0.65. These values are closer to the best results achieved with the method developed in this thesis.

Finally, in a recent study [Nair and Yu, 2020] produced a proxy for CCN at 0.4% supersaturation in SGP. They utilised machine learning, where they developed a random forest regression model based on atmospheric measurements, including PM2.5 composition fractions, trace gases, and meteorological parameters. They used Kendall rank correlation as an evaluation metric, which was around 0.53 when model-derived CCN was compared to observed CCN in SGP. Compared to the results from this thesis, the same value for the mode parameter global proxy would be 0.52, so the performance is almost same as in [Nair and Yu, 2020].

The overall conclusion from the comparisons between this thesis and previous studies is that the method developed in this thesis can achieve at best similar results but the proxy performance tends to be lower. Though, it should be noted that for some reason in SGP the proxy is not able to replicate observed N_{100} as well as in many other rural sites, so the comparison might give more favourable results if done with other sites.

The main issues with the global proxy are the inability to reproduce low and high N_{100} values. Especially the problems with predicting lower values seem to be related to how the proxy is constructed. Firstly, at cooler temperatures, the temperature-dependant part was defined by a(T<5) intercept, which in almost all cases was negative. This implies that when the temperature is low, N_{100} would be removed from the atmosphere based on the temperature-dependant part of the proxy, which is not physically accurate. The CO-dependant part then increases N_{100} based on CO concentration. For clean sites, like Alert, where the CO concentrations are low, the resulting N_{100} might still be negative so lower limit of 10 cm^{-3} was applied. On the other hand, it seems that for most sites, despite the relative error fitting method, the CO-dependant part produces too high N_{100}

at low CO concentrations. Similarly, since in reality N_{100} depends on multiple different variables, typically at certain CO concentration there is a range of possible N_{100} values that a single fit cannot capture. Therefore, in addition to better CO fitting, improving the proxy probably requires incorporating additional CAMS variables.

Additionally, data coverage hindered the analysis. One of the challenges in creating and evaluating the proxies was the varying lengths of data available for training and testing. Having several years of data from each site would allow a more confident analysis of the effect of train and test sets. Most importantly, it would be good if train and test sets would have similar distributions. The proxy parameters would account better for natural yearly variation if the global proxy could be trained on multiple years of data from each site. This would lower the risk of outlier years affecting the parameters. Finally, it would be important to have more measurement sites with sufficiently long datasets at different types of locations, especially outside Europe. The dataset used in this thesis is very eurocentric, which is likely to make the global proxy less globally representative.

7. Conclusions

In this thesis, a simple proxy for continental concentrations of accumulation mode particles was developed using reanalysis data. Proxies like this are needed to estimate global concentrations of CCN particles, which affect climate. Here, carbon monoxide (CO) and air temperature (T) were used as tracers for anthropogenic and biogenic emissions to predict the number concentration of particles with dry diameters above 100 nm (N_{100}). The dataset contained daily N_{100} measurements from 22 sites on 5 continents and was combined with CO and temperature from CAMS reanalysis. In addition to a global continental proxy, also site proxies and site excluded proxies were created.

The proxies were developed by dividing the data into temperature bins and applying N_{100} versus CO fit to each bin. To better the fit, a new fitting method was developed with an emphasis on capturing the correct order of magnitude of the accumulation mode particle concentrations. It also forced the fit to follow smaller concentrations more closely while allowing larger absolute error for higher concentrations. After fitting, the average of the bin slopes was used as CO dependant parameter b_{ave}. Temperature dependant parameter a(T) was calculated from fit intercepts so that below 5°C the temperature contribution to N_{100} was constant and above 5°C increased linearly.

Proxy evaluations were done by splitting the data into train sets, which were used to train the proxy, and test sets, where test CO and test temperature were used to calculate predicted N_{100}, which was then compared to observed N_{100}. This was repeated with the site proxy, the site excluded proxy, and the global proxy. The final global proxies were produced by selecting 100 training sets with a random year of data from each continental site, calculating the parameters, and testing against all available data from each site.

Proxy evaluation results show that proxy performance varies significantly depending on test and train sets. Some of it is caused by variation between years. For example, in Helsinki (Finland) the proxy performs better during years with warmer summers and in Alert (Canada) the CO concentrations between the two available years of data differ significantly. However, also data availability affects the variation, for instance, if sets cover only part of the year, training and testing against different seasons causes differences in performance.

The site proxy performs typically better than the site excluded proxy and the global

proxy. The sites where the proxy performs best are Mace Head (Ireland) and some of the European rural sites together with São Paulo (Brazil). Mountainous sites Hohenpeissenberg and Schauinsland (Germany) have poor results, which is likely related to boundary layer height. Additionally, Southern Great Planes (USA) and K-Puszta (Hungary) perform worse compared to other sites, and especially in K-Puszta proxies trained with other sites' data cannot predict N_{100} properly.

For the final global proxy, selecting one parameter set that would perform best for all sites or certain site types was not possible. Instead, two parameter sets were selected and the results from them investigated. For Marikana (South Africa), Nanjing (China), and São Paulo (Brazil) the difference in parameter sets does not affect the results, possibly because in polluted urban sites the proxy follows anthropogenic emissions and CO dependant parameter b_{ave} varies proportionally less than parameter a(T). K-Puszta performs better with a different parameter set compared to all other sites. However, with this set the proxy performance in Alert (Canada) decreases significantly and results in extreme overestimation of N_{100}. Overall, the global proxy typically underestimates high concentrations, overestimates low concentrations, or both. This varies depending on the season, but generally, the method developed in this thesis struggles to capture the entire range of observed N_{100}. Particularly concentrations below 500 cm^{-3} are challenging. As a result, the global proxy is not able to predict N_{100} with sufficient accuracy. Especially remote and clean sites would require good accuracy, but the global proxy can not capture low enough concentrations reliably to achieve this. When comparing the results to CCN proxies from literature, the method developed in this thesis performs typically less well or at best as well as other methods. However, it should be noted that the method in this thesis is quite simple with only two variables used in the proxy. Additionally, since the proxy performance depends strongly on the site, a better comparison to literature would require more studies that have results for same sites as this thesis.

Another limitation of the developed proxy is its applicability in locations that do not have reference particle measurements. Currently, extrapolating the global proxy to these sites would produce results with highly varying accuracy. Therefore, currently, the proxy cannot be used to reliably predict N_{100} outside measurement sites.

In the future, the proxy could be further developed with larger datasets that contain at least two years of data from each site. This would already reduce the uncertainties during proxy evaluation, though having several years of simultaneous data from all sites would be ideal. Additionally, increasing the number of sites outside Europe could improve proxy's performance globally, or at least help evaluating the accuracy. With additional sites, it could also be possible to find site types where the proxy performs well. Finally, some additional variables, like boundary layer height, could help the proxy predict lower and higher N_{100} concentrations.

8. Acknowledgements

The proxy was generated using Copernicus Atmosphere Monitoring Service information 2021. CAMS reanalysis dataset was downloaded from the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS) `https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview`

Appendix A. Predicted and Observed Time Series of N_{100} for the Site Proxy

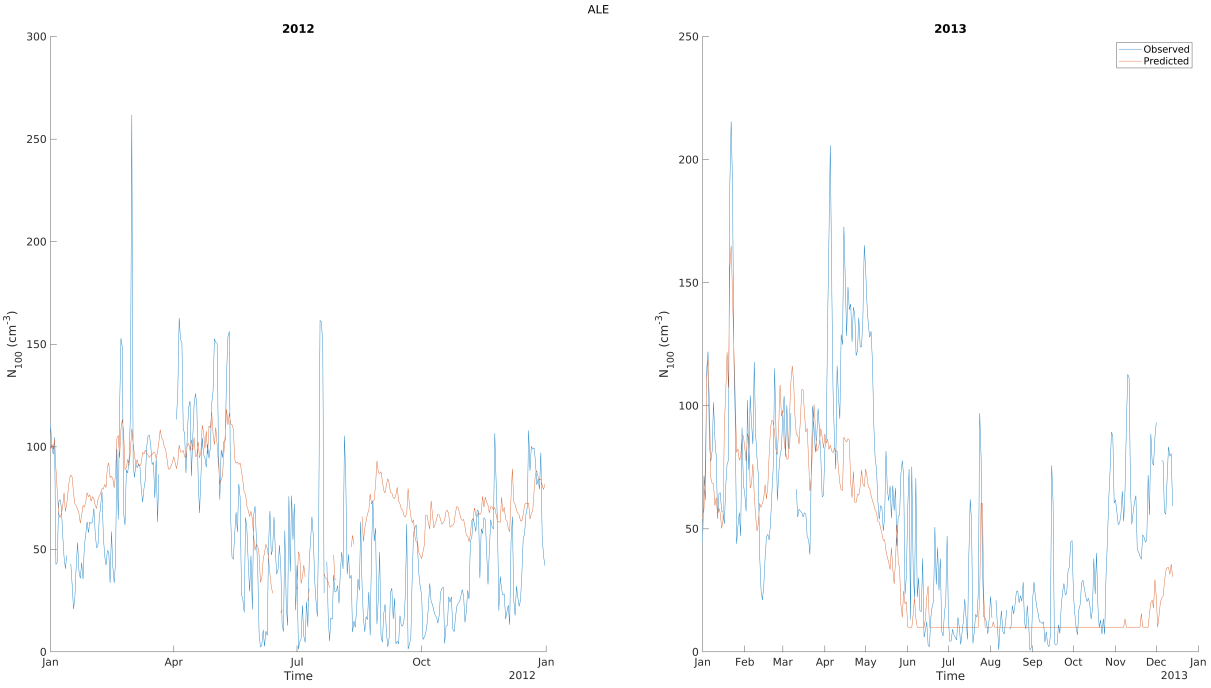

Figure A.1: Time series of observed and predicted N_{100} for Alert (ALE). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

Figure A.2: Time series of observed and predicted N_{100} for Annaberg-Buchholz (ABZ). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

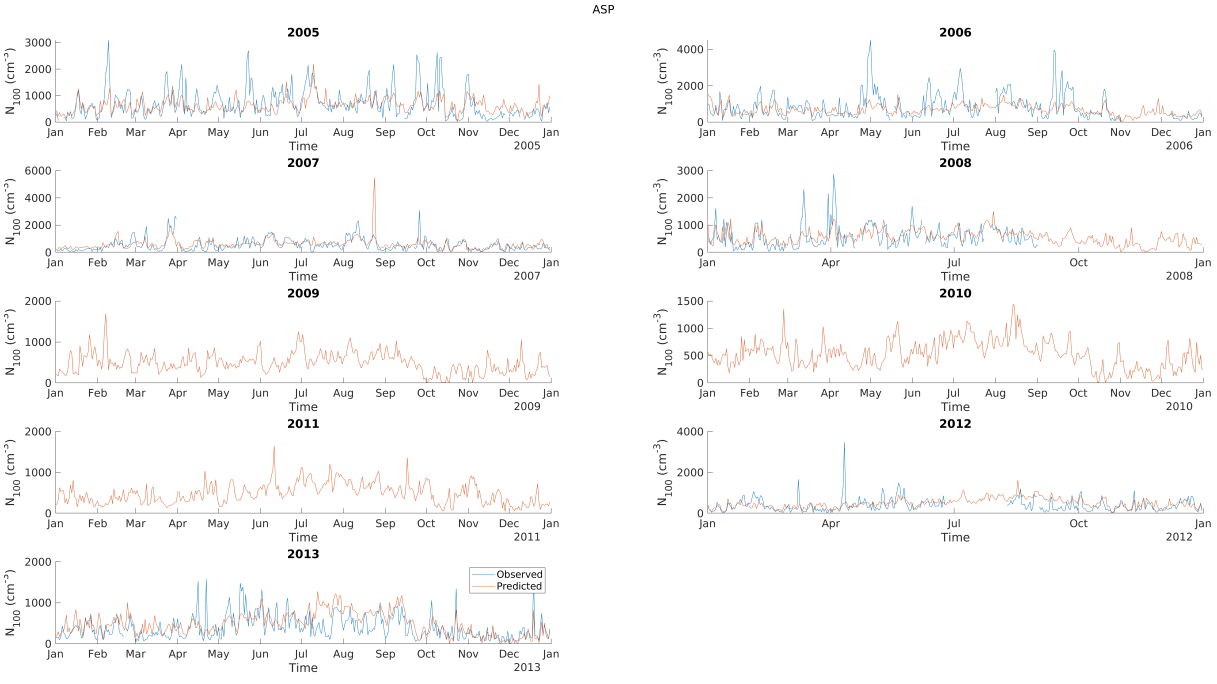

Figure A.3: Time series of observed and predicted N_{100} for Aspverten (ASP). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

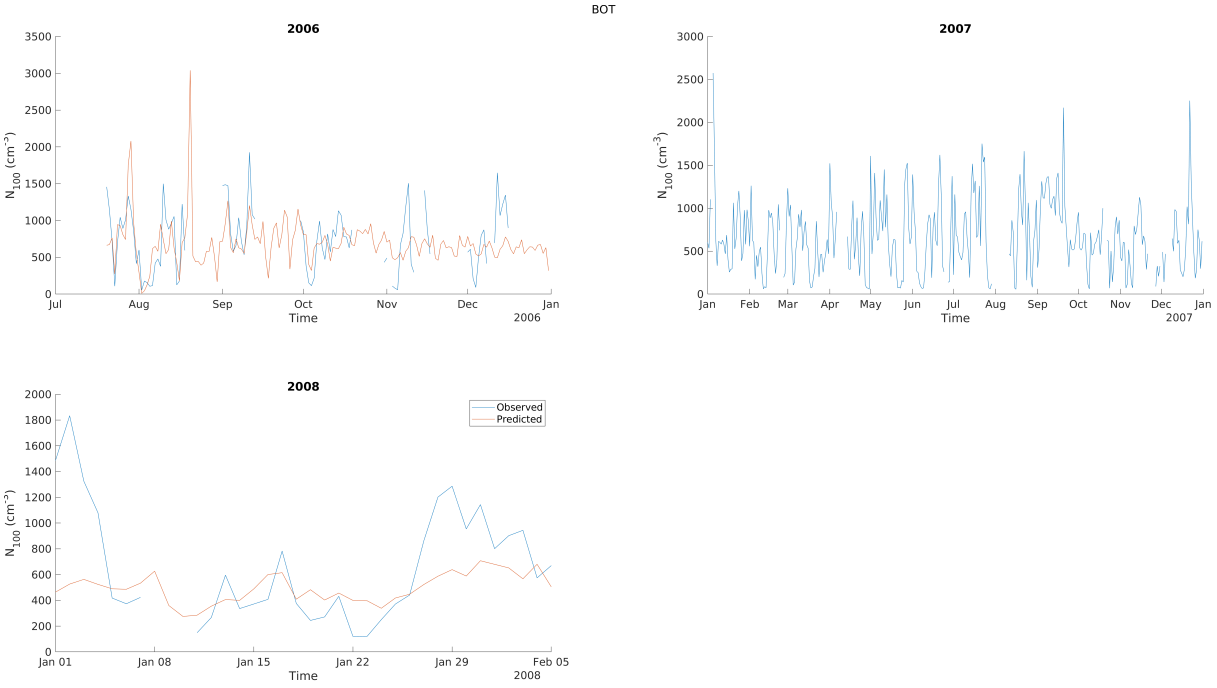

Figure A.4: Time series of observed and predicted N_{100} for Botsalano (BOT). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

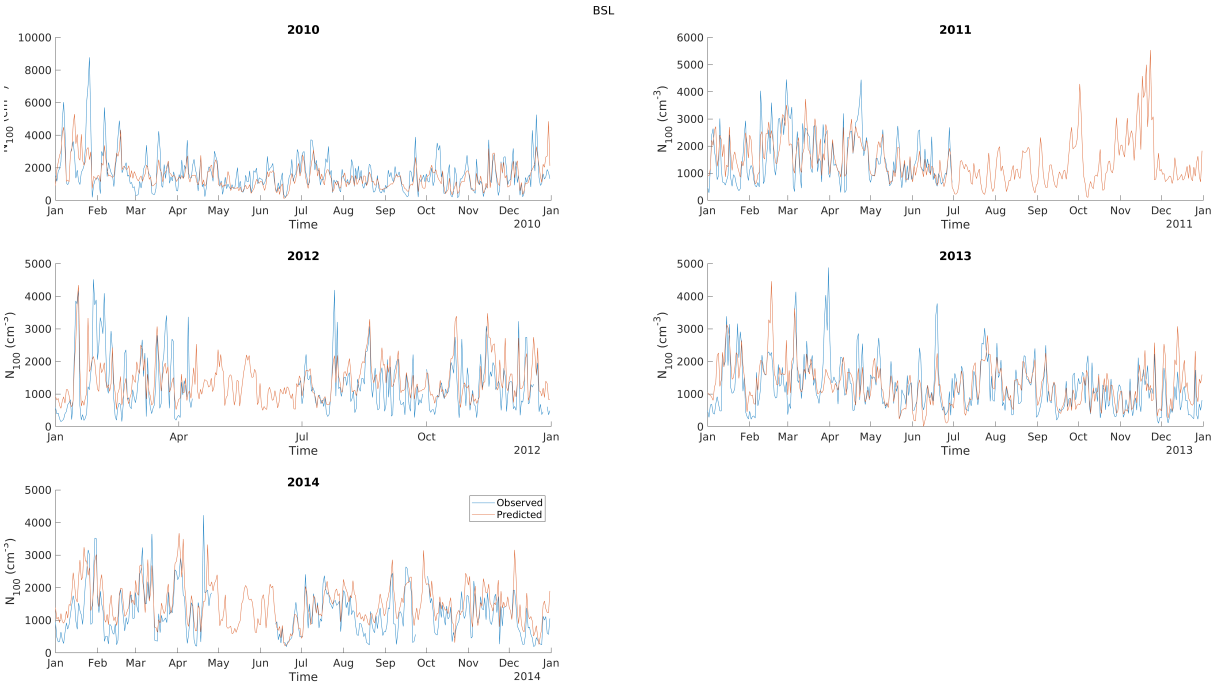

Figure A.5: Time series of observed and predicted N_{100} for Bösel (BSL). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

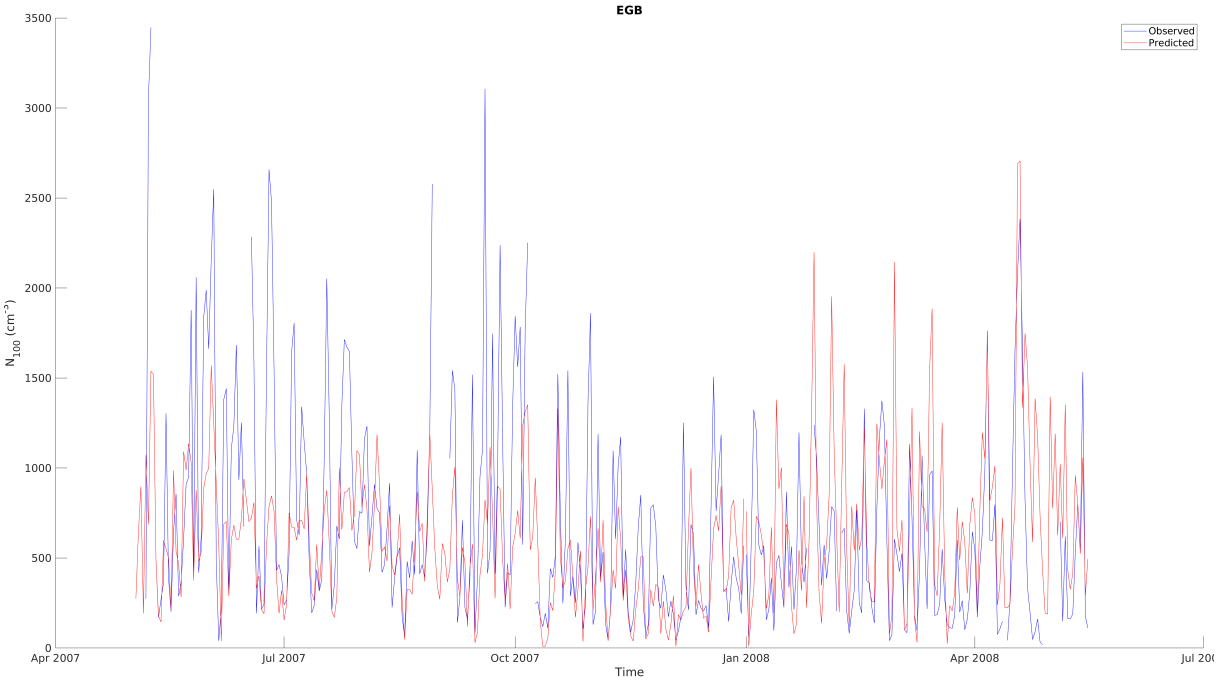

Figure A.6: Time series of observed and predicted N_{100} for Egbert (EGB). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

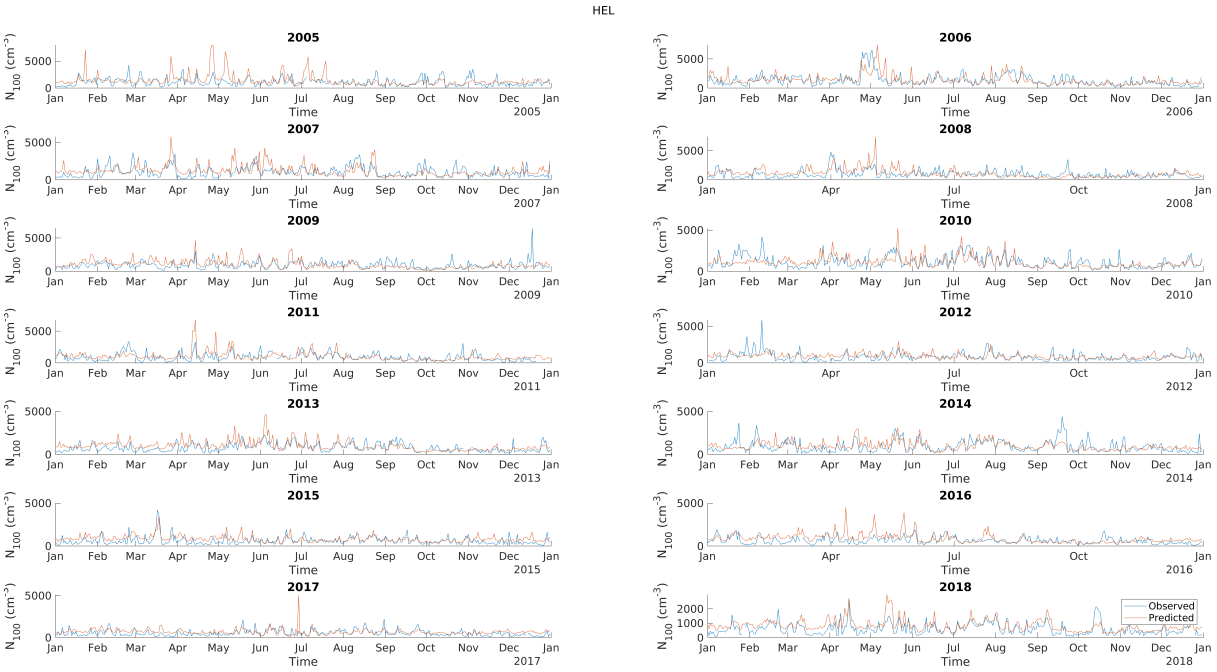

Figure A.7: Time series of observed and predicted N_{100} for Helsinki (HEL). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

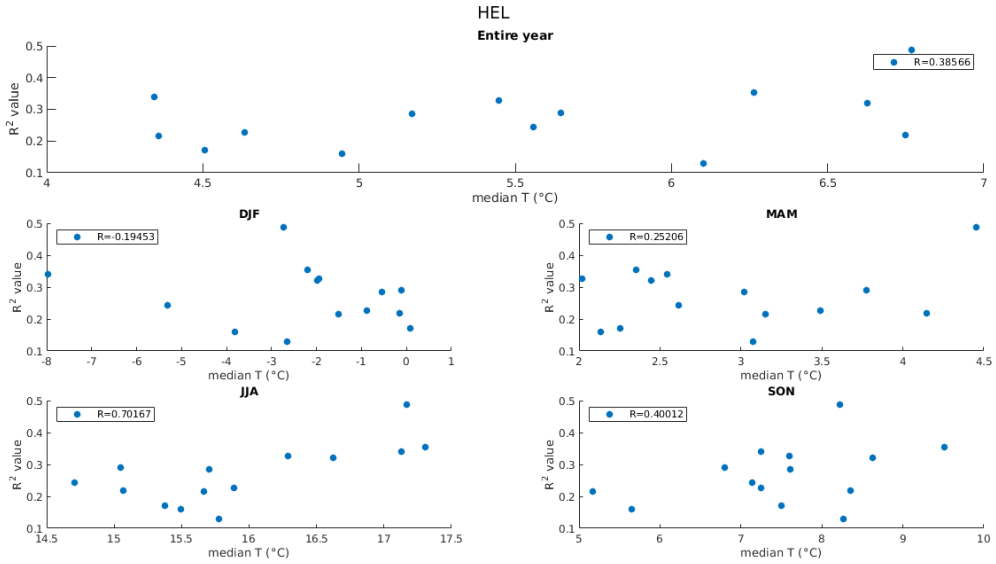

Figure A.8: Figure shows the proxy R^2 in different years against median temperatures, with upper panel showing R^2 against annual median temperature and lower panels the seasonal median temperatures. R shows the Pearson correlation coefficient between R^2 values and median temperatures.

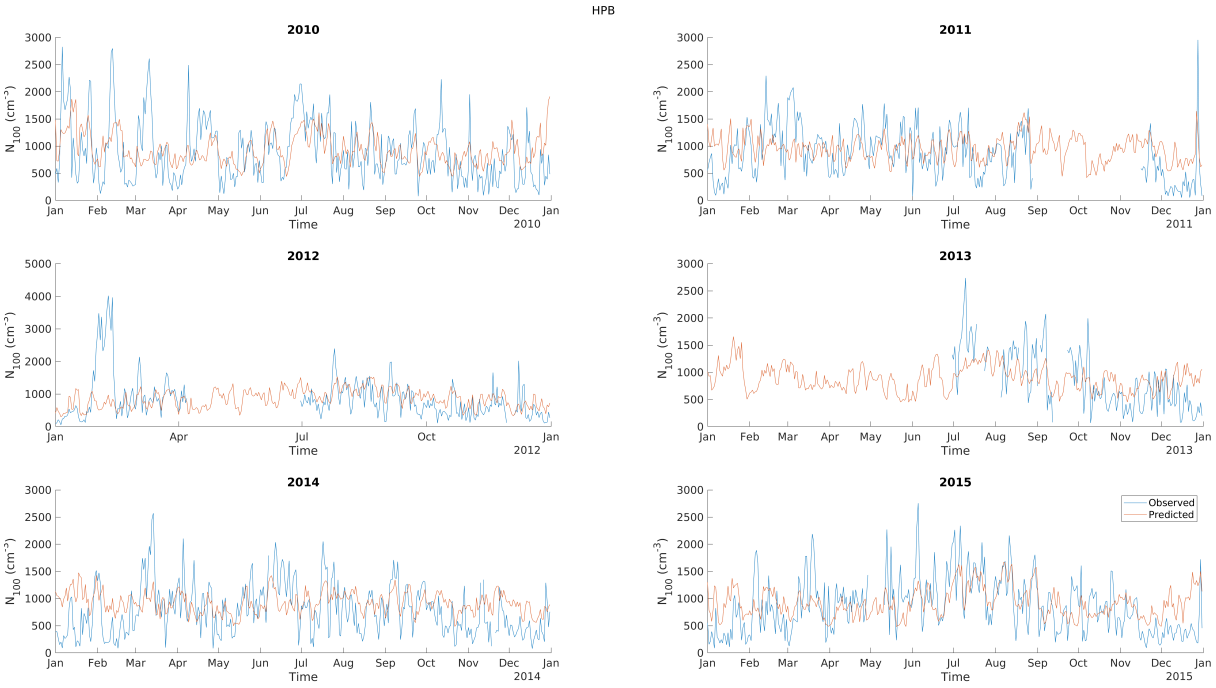

Figure A.9: Time series of observed and predicted N_{100} for Hohenpeissenberg (HPB). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

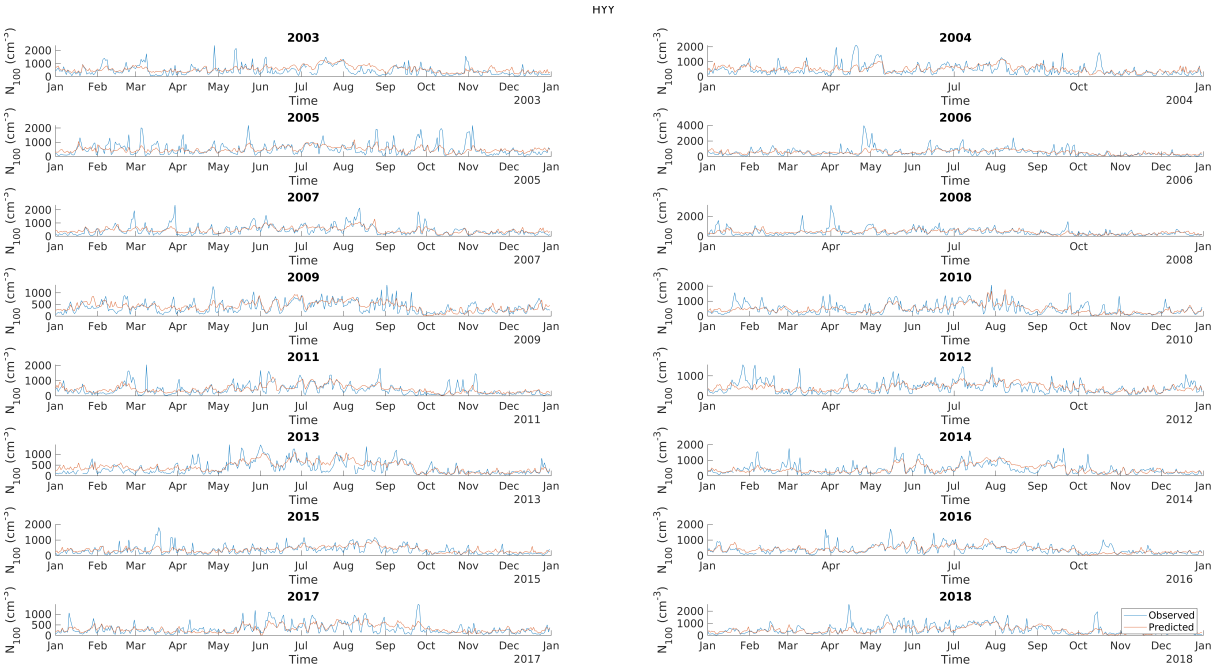

Figure A.10: Time series of observed and predicted N_{100} for Hyytiälä (HYY). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

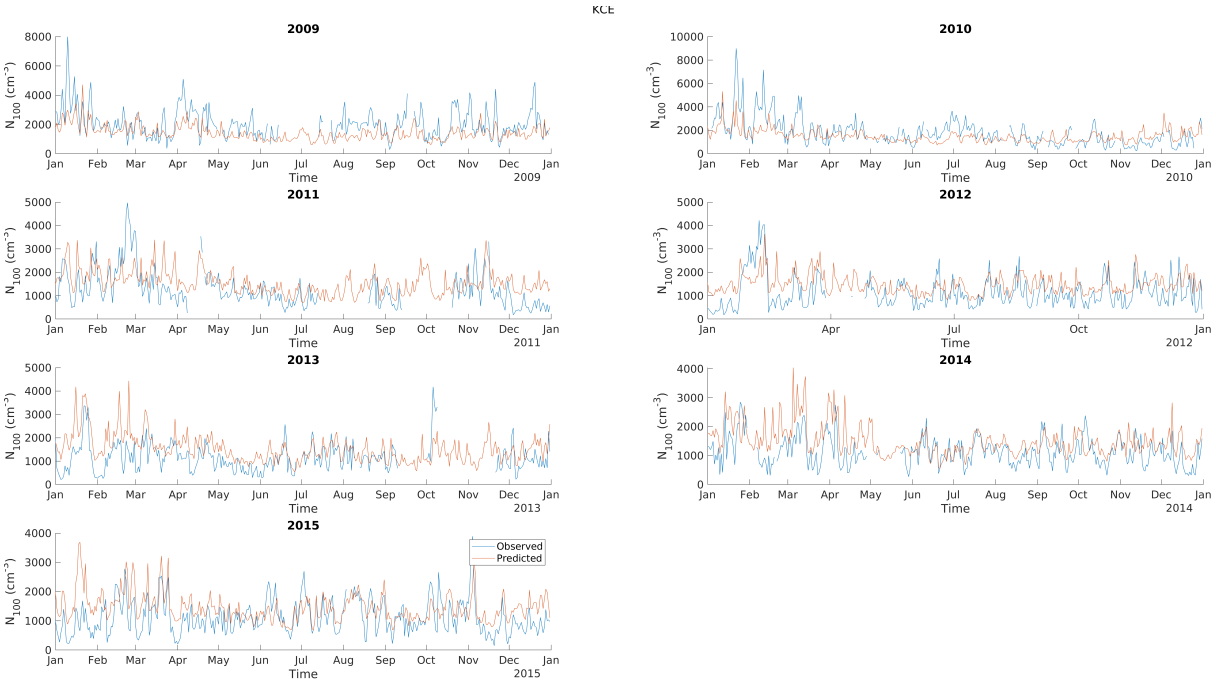

Figure A.11: Time series of observed and predicted N_{100} for Kosetice (KCE). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

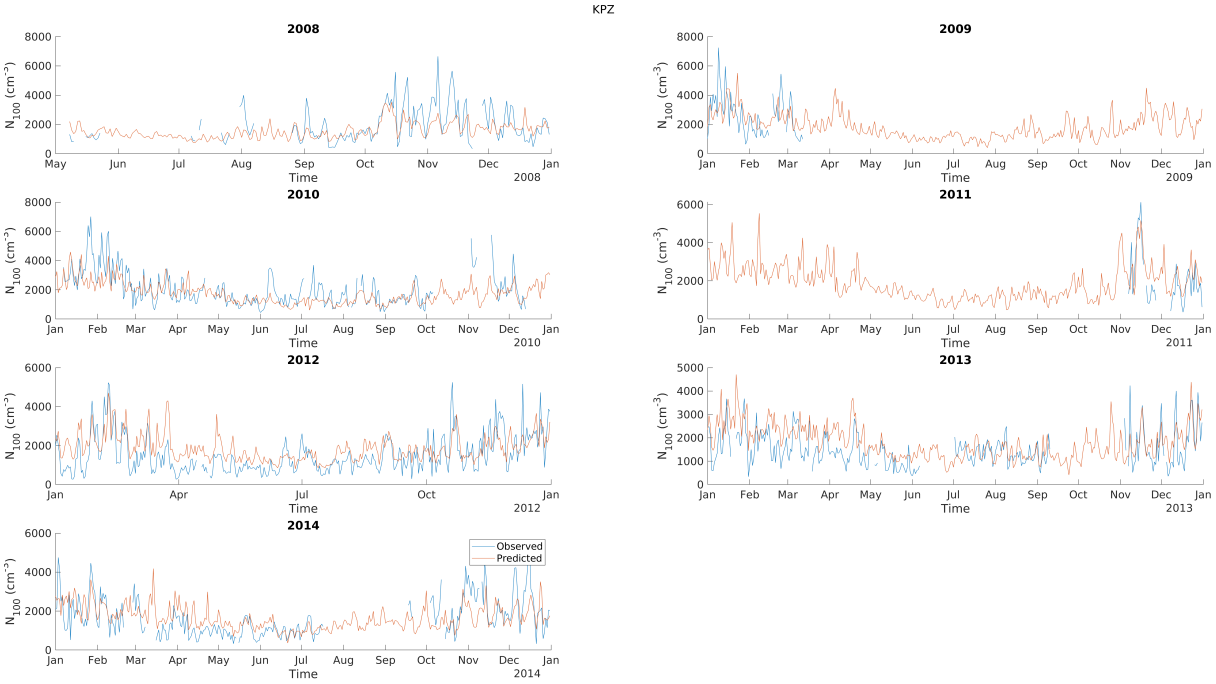

Figure A.12: Time series of observed and predicted N_{100} for K-Puszta (KPZ). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

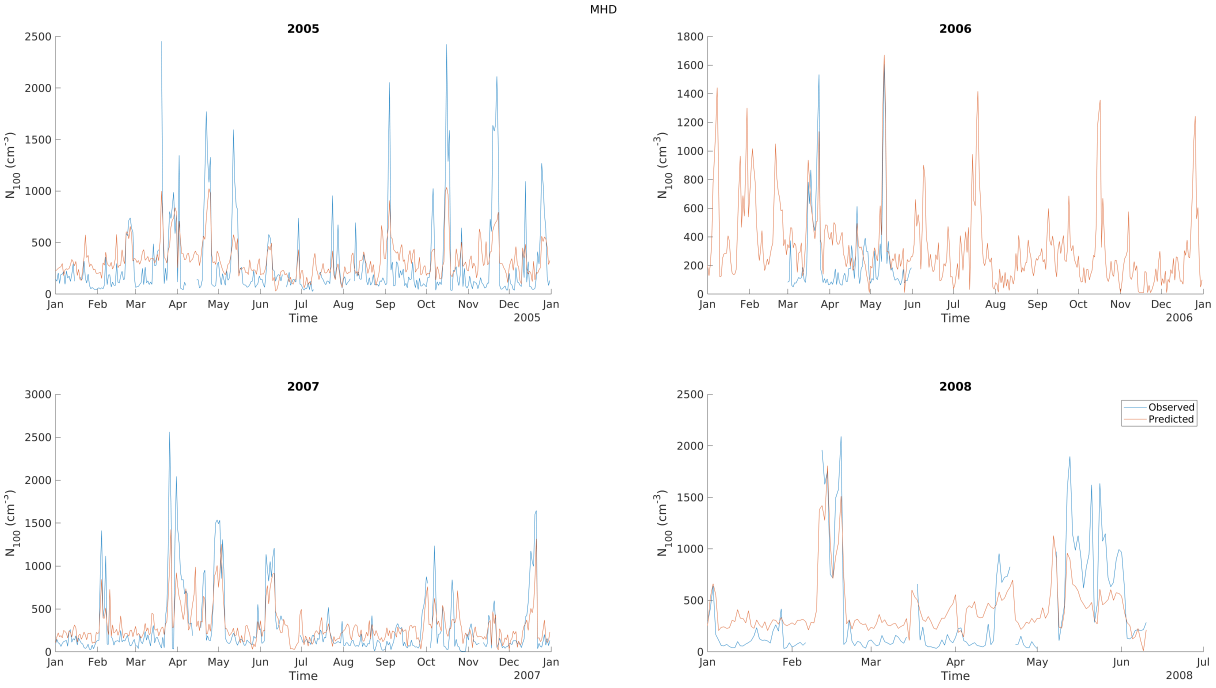

Figure A.13: Time series of observed and predicted N_{100} for Mace Head (MHD). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

Figure A.14: Time series of observed and predicted N_{100} for Marikana (MAR). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

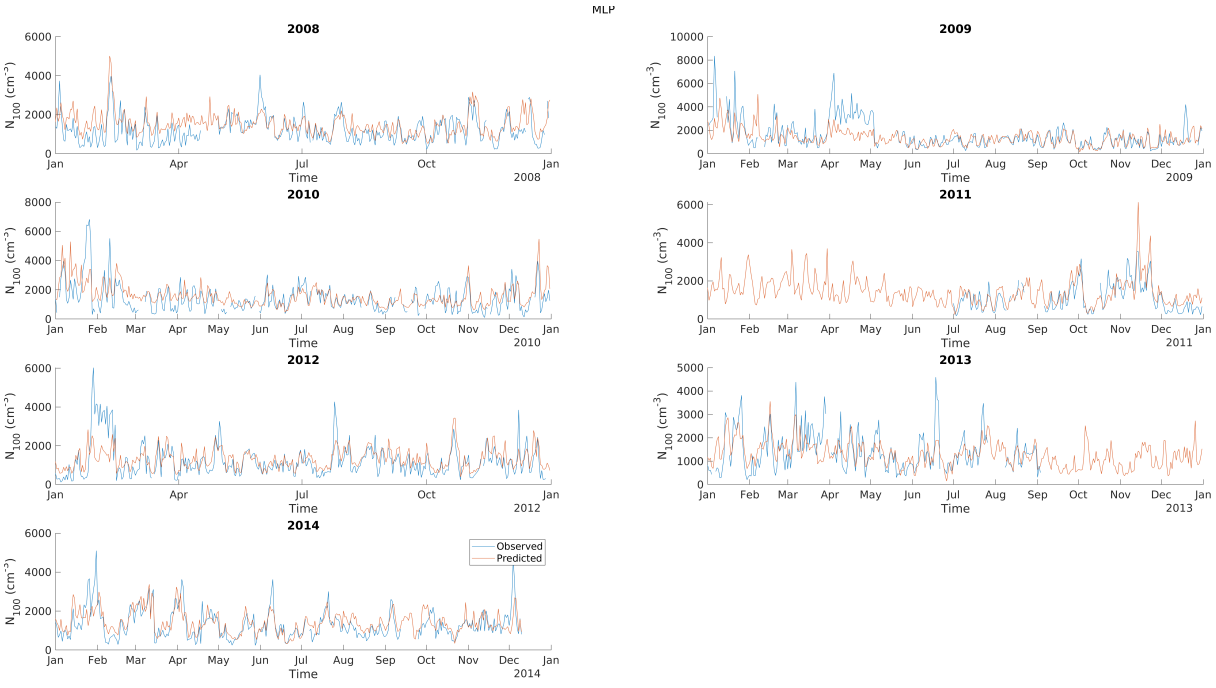

Figure A.15: Time series of observed and predicted N_{100} for Melpitz (MLP). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

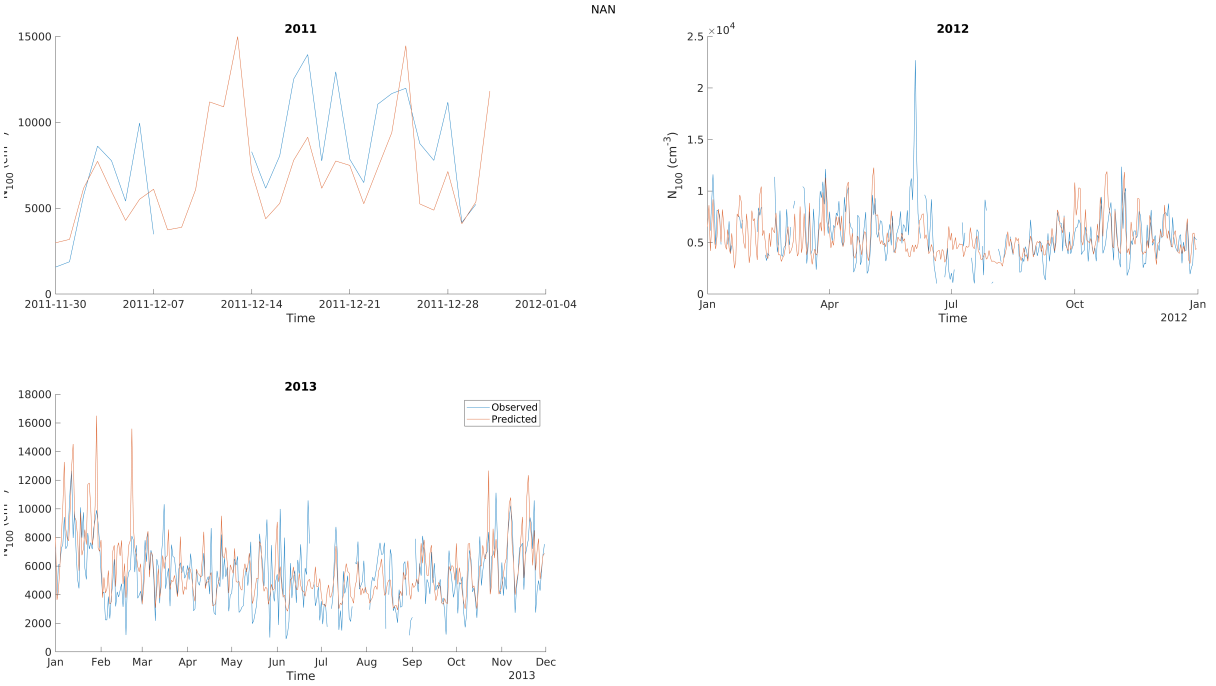

Figure A.16: Time series of observed and predicted N_{100} for Nanjing (NAN). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

Figure A.17: Time series of observed and predicted N_{100} for Neuglobsow (NEU). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

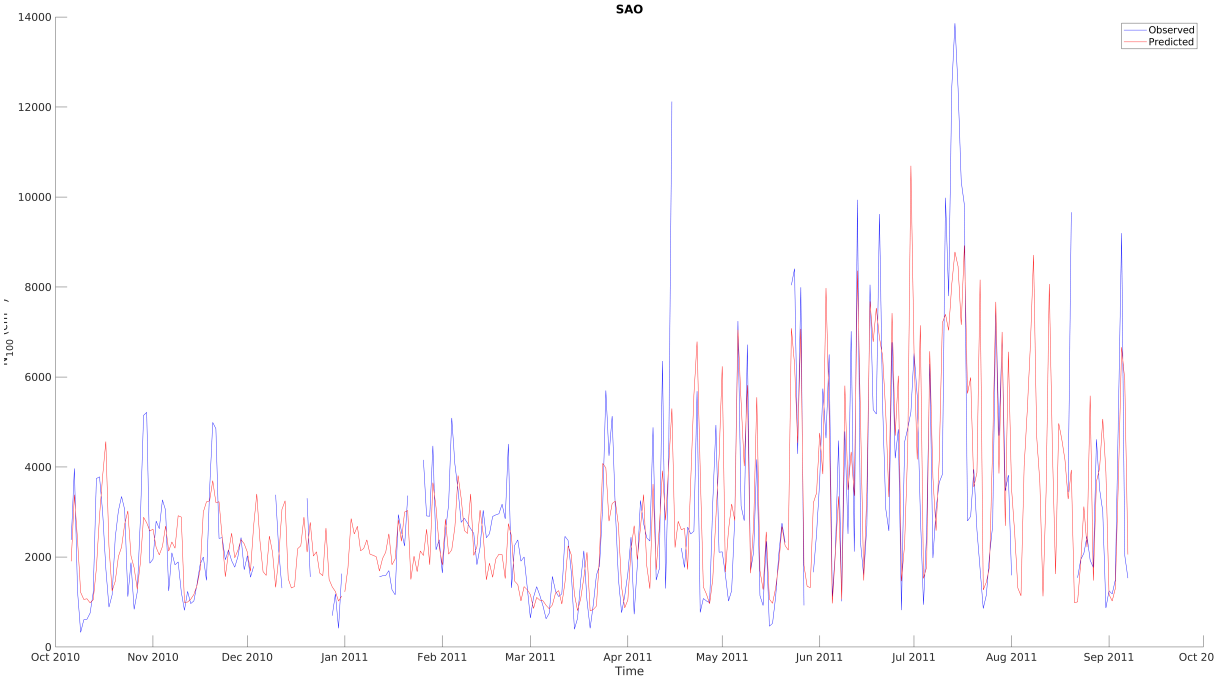

Figure A.18: Time series of observed and predicted N_{100} for Sao Paulo (SAO). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

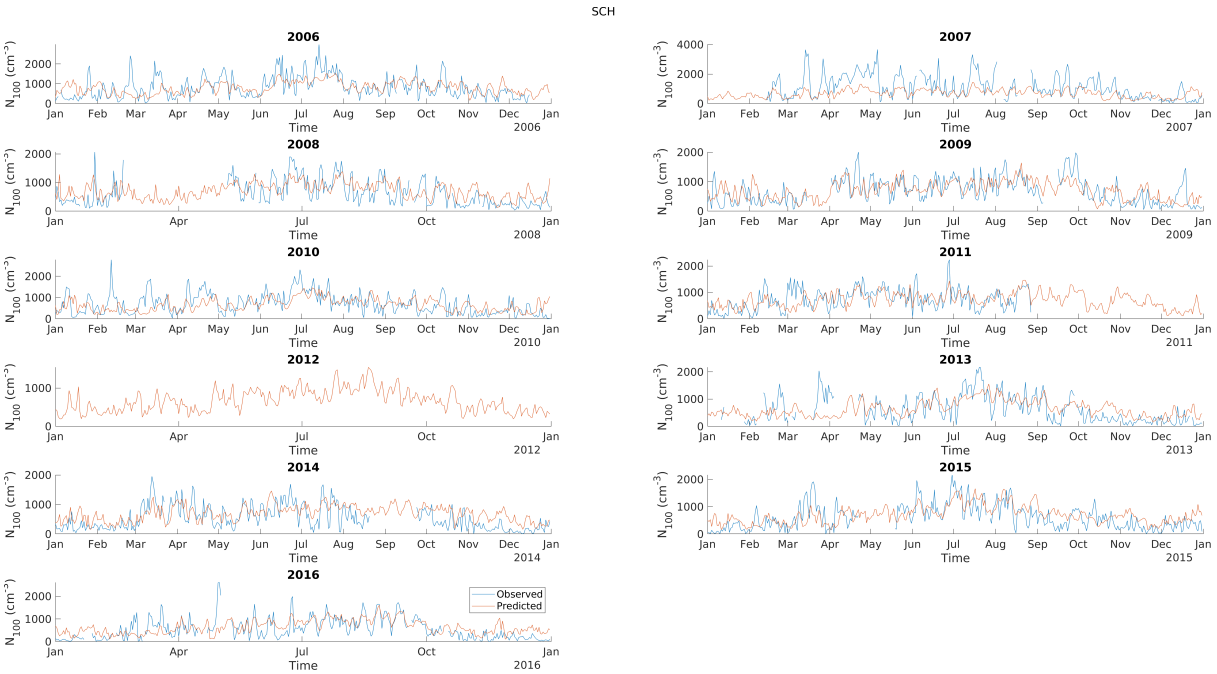

Figure A.19: Time series of observed and predicted N_{100} for Schauinsland (SCH). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

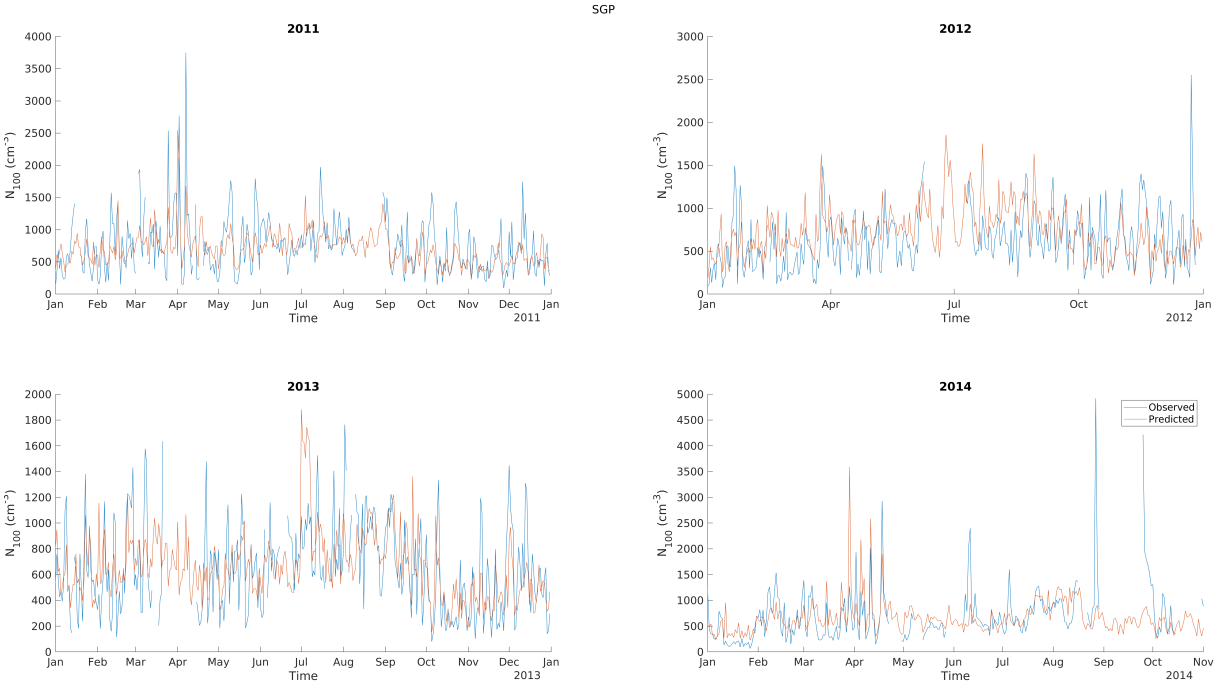

Figure A.20: Time series of observed and predicted N_{100} for Southern Great Planes (SGP). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

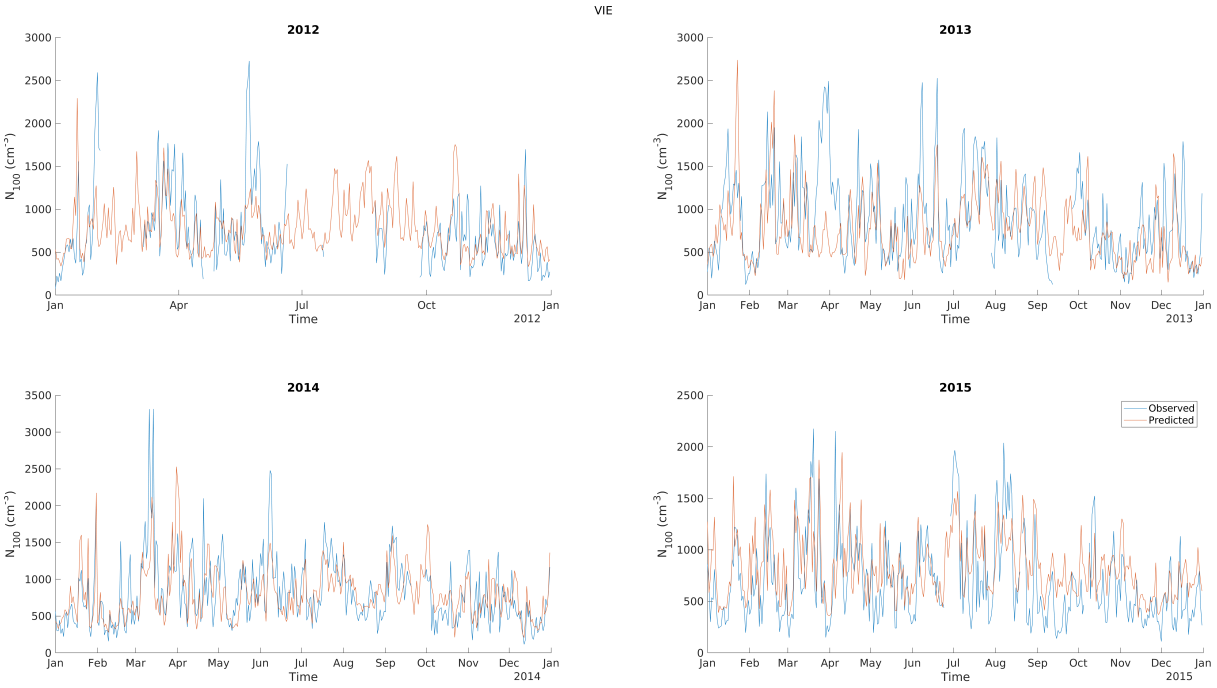

Figure A.21: Time series of observed and predicted N_{100} for Vielsalm (VIE). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

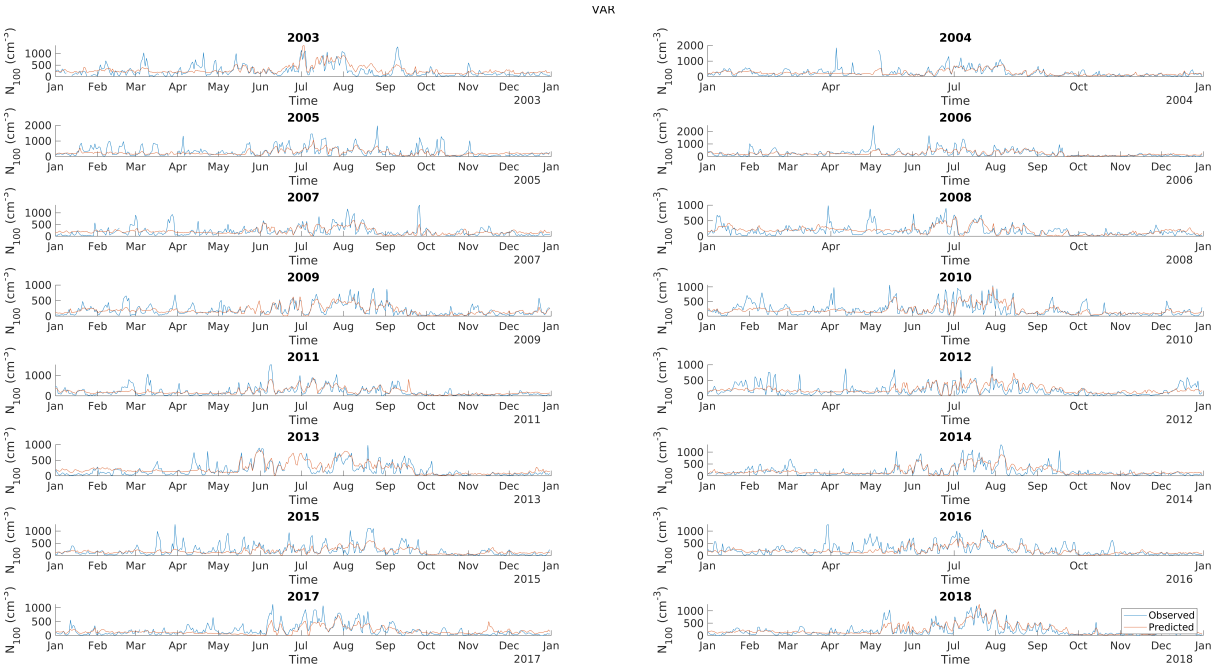

Figure A.22: Time series of observed and predicted N_{100} for Värriö (VAR). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

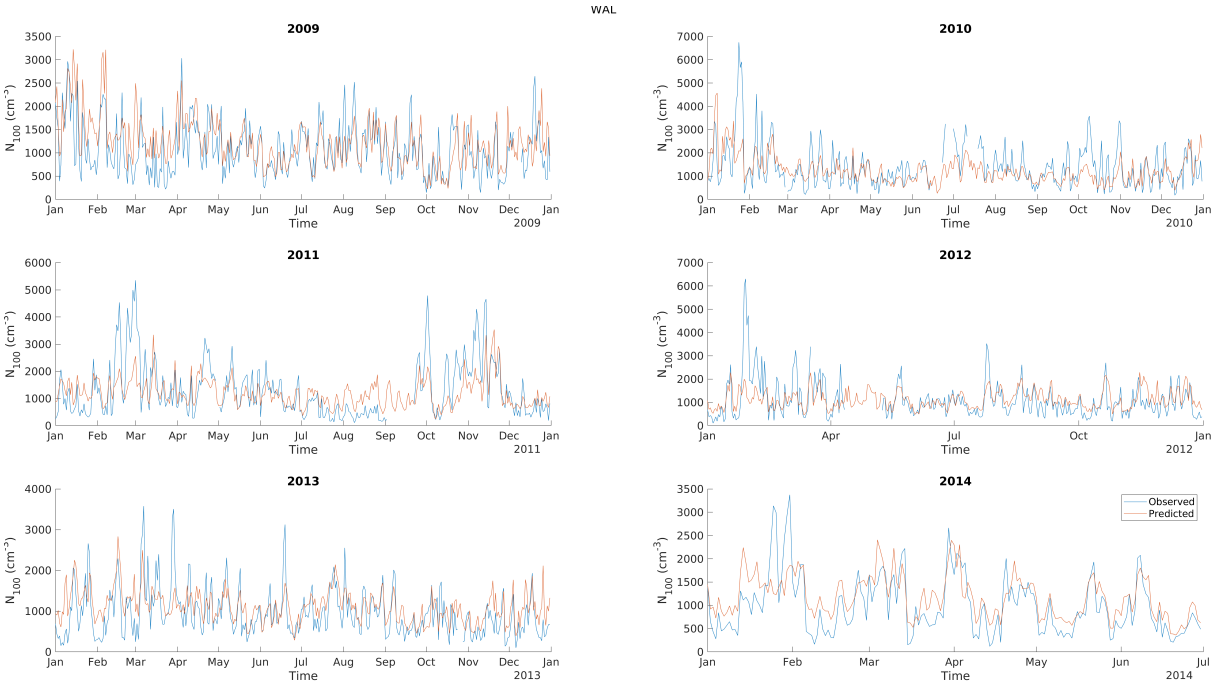

Figure A.23: Time series of observed and predicted N_{100} for Waldhof (WAL). Proxy was trained with site's own data so that target year's data was left out when calculating proxy parameters, and then the predicted N_{100} was calculated with parameters along with temperature and CO data from target year.

Bibliography

[Aalto et al., 2001] Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mäkelä, J. M., Hoell, C., Ó dowd, C. D., Hansson, H.-C., Väkevä, M., Koponen, I. K., Buzorius, G., and Kulmala, M. (2001). Physical characterization of aerosol particles during nucleation events. *Tellus. Series B: Chemical and Physical Meteorology*, 53(4):344–358.

[Andreae and Rosenfeld, 2008] Andreae, M. O. and Rosenfeld, D. (2008). Aerosol-cloud-precipitation interactions. part 1. the nature and sources of cloud-active aerosols. *Earth-Science Reviews*, 89(1):13–41.

[Bergmeir and Benítez, 2012] Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

[Birmili et al., 2016] Birmili, W., Weinhold, K., Rasch, F., Sonntag, A., Sun, J., Merkel, M., Wiedensohler, A., Bastian, S., Schladitz, A., Löschau, G., Cyrys, J., Pitz, M., Gu, J., Kusch, T., Flentje, H., Quass, U., Kaminski, H., Kuhlbusch, T. A. J., Meinhardt, F., Schwerin, A., Bath, O., Ries, L., Wirtz, K., and Fiebig, M. (2016). Long-term observations of tropospheric particle number size distributions and equivalent black carbon mass concentrations in the german ultrafine aerosol network (guan). *Earth system science data*, 8(2):355–382.

[Boucher et al., 2013] Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., and Zhang, X. (2013). *Clouds and Aerosols*, book section 7, pages 571–658. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[Després et al., 2012] Després, V., Huffman, J. A., Burrows, S. M., Hoose, C., Safatov, A., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M., Pöschl, U., and Jaenicke, R. (2012). Primary biological aerosol particles in the atmosphere: a review. *Tellus B: Chemical and Physical Meteorology*, 64(1):15598. doi: 10.3402/tellusb.v64i0.15598.

[Dusek et al., 2006] Dusek, U., Frank, G. P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., Chand, D., Drewnick, F., Hings, S., Jung, D., Borrmann, S., and Andreae,

M. O. (2006). Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science (American Association for the Advancement of Science); Science*, 312(5778):1375–1378.

[Fays et al., 2019] Fays, S., Laruelle, R., Luthers, C., and Gérard, G. (2019). Ultrafine particle measurements in wallonia, belgium. *Congrés Français sur les Aérosols 2019, Paris*.

[Gentner et al., 2017] Gentner, D. R., Jathar, S. H., Gordon, T. D., Bahreini, R., Day, D. A., Haddad, I. E., Hayes, P. L., Pieber, S. M., Platt, S. M., de Gouw, J., Goldstein, A. H., Harley, R. A., Jimenez, J. L., H., A. S. P., and Robinson, A. L. (2017). Review of urban secondary organic aerosol formation from gasoline and diesel motor vehicle emissions. *Environmental science & technology; Environ.Sci.Technol*, 51(3):1074–1093.

[Guyon et al., 2005] Guyon, P., Frank, G. P., Welling, M., Chand, D., Artaxo, P., Rizzo, L., Nishioka, G., Kolle, O., Fritsch, H., F., M. A. S. D., Gatti, L. V., Cordova, A. M., and Andreae, M. O. (2005). Airborne measurements of trace gas and aerosol particle emissions from biomass burning in amazonia. *Atmospheric chemistry and physics*, 5(11):2989–3002.

[Inness et al., 2019a] Inness, A., Ades, M., Agusti-Panareda, A., Barre, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M. (2019a). Cams global reanalysis (eac4). copernicus atmosphere monitoring service (cams) atmosphere data store (ads). (Accessed on 05-04-2021).

[Inness et al., 2019b] Inness, A., Ades, M., Agusti-Panareda, A., Barre, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M. (2019b). The cams reanalysis of atmospheric composition. *Atmospheric chemistry and physics*, 19(6):3515–3556.

[Liu and Li, 2014] Liu, J. and Li, Z. (2014). Estimation of cloud condensation nuclei concentration from aerosol optical quantities: influential factors and uncertainties. *Atmospheric chemistry and physics*, 14(1):471–483.

[McFiggans et al., 2006] McFiggans, G., Artaxo, P., Baltensperger, U., Coe, H., Facchini, M. C., Feingold, G., Fuzzi, S., Gysel, M., Laaksonen, A., Lohmann, U., Mentel, T. F., Murphy, D. M., D., C. O., Snider, J. R., and Weingartner, E. (2006). The effect of physical and chemical aerosol properties on warm cloud droplet activation. *Atmospheric chemistry and physics*, 6(9):2593–2649.

[Merikanto et al., 2009] Merikanto, J., Spracklen, D. V., Mann, G. W., Pickering, S. J., and Carslaw, K. S. (2009). Impact of nucleation on global ccn. *Atmospheric chemistry and physics*, 9(21):8601–8616.

[Nair and Yu, 2020] Nair, A. A. and Yu, F. (2020). Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric chemistry and physics*, 20(21):12853–12869.

[Nieminen et al., 2018] Nieminen, T., Kerminen, V.-M., Petaja, T., Aalto, P. P., Arshinov, M., Asmi, E., Baltensperger, U., Beddows, D. C. S., Beukes, J. P., Collins, D., Ding, A., Harrison, R. M., Henzing, B., Hooda, R., Hu, M., Horrak, U., Kivekas, N., Komsaare, K., Krejci, R., Kristensson, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Lihavainen, H., Mihalopoulos, N., Nemeth, Z., Nie, W., O'Dowd, C., Salma, I., Sellegri, K., Svenningsson, B., Swietlicki, E., Tunved, P., Ulevicius, V., Vakkari, V., Vana, M., Wiedensohler, A., Wu, Z., Virtanen, A., and Kulmala, M. (2018). Global analysis of continental boundary layer new particle formation based on long-term measurements. *Atmospheric chemistry and physics*, 18(19):14737–14756.

[Paasonen et al., 2013] Paasonen, P., Asmi, A., Petäjä, T., Kajos, M., Äijälä, M., Junninen, H., Holst, T., Abbatt, J. P. D., Arneth, A., Birmili, W., Van Der Gon, H. D., Hamed, A., Hoffer, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Plass-dülmer, C., Pryor, S. C., Räisänen, P., Swietlicki, E., Wiedensohler, A., Worsnop, D. R., Kerminen, V.-M., and Kulmala, M. (2013). Warming-induced increase in aerosol number concentration likely to moderate climate change. *Nature Geoscience*, 6(6):438–442.

[Paramonov et al., 2015] Paramonov, M., Kerminen, V.-M., Gysel, M., Aalto, P. P., Andreae, M. O., Asmi, E., Baltensperger, U., Bougiatioti, A., Brus, D., Frank, G. P., Good, N., Gunthe, S. S., Hao, L., Irwin, M., Jaatinen, A., JurÃ¡nyi, Z., King, S. M., Kortelainen, A., Kristensson, A., Lihavainen, H., Kulmala, M., Lohmann, U., Martin, S. T., McFiggans, G., Mihalopoulos, N., Nenes, A., O'Dowd, C. D., Ovadnevaite, J., Petäjä, T., Pöschl, U., Roberts, G. C., Rose, D., Svenningsson, B., Swietlicki, E., Weingartner, E., Whitehead, J., Wiedensohler, A., Wittbom, C., and Sierau, B. (2015). A synthesis of cloud condensation nuclei counter (ccnc) measurements within the eucaari network. *Atmospheric chemistry and physics*, 15(21):12211–12229.

[Petters and Kreidenweis, 2007] Petters, M. D. and Kreidenweis, S. M. (2007). A single parameter representation of hygroscopic growth and cloud condensation nucleus activity. *Atmospheric chemistry and physics*, 7(8):1961–1971.

[Pierce and Adams, 2008] Pierce, J. R. and Adams, P. J. (2008). Uncertainty in global ccn concentrations from uncertain aerosol nucleation and primary emission rates. *Atmospheric chemistry and physics discussions*, 8(4):16291–16333.

[Reid et al., 2005] Reid, J. S., Koppmann, R., Eck, T. F., and Eleuterio, D. P. (2005). A review of biomass burning emissions part ii: intensive physical properties of biomass burning particles. *Atmospheric chemistry and physics*, 5(3):799–825.

[Riipinen et al., 2011] Riipinen, I., Pierce, J. R., Yli-Juuti, T., Nieminen, T., Häkkinen, S., Ehn, M., Junninen, H., Lehtipalo, K., Petäjä, T., Slowik, J., Chang, R., Shantz, N. C., Abbatt, J., Leaitch, W. R., Kerminen, V. M., Worsnop, D. R., Pandis, S. N., Donahue, N. M., and Kulmala, M. (2011). Organic condensation: a vital link connecting aerosol formation to cloud condensation nuclei (ccn) concentrations. *Atmospheric chemistry and physics*, 11(8):3865–3878.

[Rosenfeld et al., 2014] Rosenfeld, D., Andreae, M. O., Asmi, A., Chin, M., de Leeuw, G., Donovan, D. P., Kahn, R., Kinne, S., Kivekäs, N., Kulmala, M., Lau, W., Schmidt, K. S., Suni, T., Wagner, T., Wild, M., and Quaas, J. (2014). Global observations of aerosol-cloud-precipitation-climate interactions: Aerosol-cloud-climate interactions. *Reviews of geophysics (1985)*, 52(4):750–808.

[Schmale et al., 2018] Schmale, J., Henning, S., Decesari, S., Henzing, B., Keskinen, H., Sellegri, K., Ovadnevaite, J., Pöhlker, M., Brito, J., Bougiatioti, A., Kristensson, A., Kalivitis, N., Stavroulas, I., Carbone, S., Jefferson, A., Park, M., Schlag, P., Iwamoto, Y., Aalto, P., Äijälä, M., Bukowiecki, N., Ehn, M., Fröhlich, R., Frumau, A., Herrmann, E., Herrmann, H., Holzinger, R., Kos, G., Kulmala, M., Mihalopoulos, N., Nenes, A., O'Dowd, C., Petäjä, T., Picard, D., Pöhlker, C., Pöschl, U., Poulain, L., Swietlicki, E., Aneae, M., Artaxo, P., Wiedensohler, A., Ogren, J., Matsuki, A., Yum, S. S., Stratmann, F., Baltensperger, U., and Gysel, M. (2018). Long-term cloud condensation nuclei number concentration, particle number size distribution and chemical composition measurements at regionally representative observatories. *Atmospheric chemistry and physics*, 18(4):2853–2881.

[Seinfeld and Pandis, 2016] Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, Incorporated, New York.

[Shen et al., 2019] Shen, Y., Virkkula, A., Ding, A., Luoma, K., Keskinen, H., Aalto, P. P., Chi, X., Qi, X., Nie, W., Huang, X., Petäjä, T., Kulmala, M., and Kerminen, V.-M. (2019). Estimating cloud condensation nuclei number concentrations using aerosol

optical properties: role of particle number size distribution and parameterization. *Atmospheric chemistry and physics*, 19(24):15483–15502.

[Shrivastava et al., 2017] Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petäjä, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q. (2017). Recent advances in understanding secondary organic aerosol: Implications for global climate forcing. *Reviews of geophysics (1985)*, 55(2):509–559.

[Small et al., 2009] Small, J. D., Chuang, P. Y., Feingold, G., and Jiang, H. (2009). Can aerosol decrease cloud lifetime? *Geophysical Research Letters*, 36(16). https://doi.org/10.1029/2009GL038888; 18.

[Spracklen et al., 2011] Spracklen, D. V., Carslaw, K. S., Pöschl, U., Rap, A., and Forster, P. M. (2011). Global cloud condensation nuclei influenced by carbonaceous combustion aerosol. *Atmospheric chemistry and physics*, 11(17):9067–9087.

[Stier, 2016] Stier, P. (2016). Limitations of passive remote sensing to constrain global cloud condensation nuclei. *Atmospheric chemistry and physics*, 16(10):6595–6607.

[Toll et al., 2019] Toll, V., Christensen, M., Quaas, J., and Bellouin, N. (2019). Weak average liquid-cloud-water response to anthropogenic aerosols. *Nature (London); Nature*, 572(7767):51–55.

[Tunved and Ström, 2019] Tunved, P. and Ström, J. (2019). On the seasonal variation in observed size distributions in northern europe and their changes with decreasing anthropogenic emissions in europe: climatology and trend analysis based on 17 years of data from aspvreten, sweden. *Atmospheric chemistry and physics*, 19(23):14849–14873.

[Twomey, 1977] Twomey, S. (1977). The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences*, 34(7):1149–1152.

[Wiedensohler et al., 2012] Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., M., A. F., Asmi, E., Sellegri, K., Depuy, R., Venzac, H., Villani, P., Laj, P., Aalto, P., Ogren, J. A., Swietlicki, E., Williams, P., Roldin, P., Quincey, P., Hüglin, C., Fierz-Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Grüning, C., Faloon, K., Beddows, D., Harrison, R., Monahan, C., Jennings, S. G., D., C. O., Marinoni, A., Horn, H. G., Keck, L., Jiang, J., Scheckman, J., McMurry, P. H., Deng, Z., Zhao, C. S., Moerman, M., Henzing, B., de Leeuw, G., Löschau, G., and Bastian, S. (2012). Mobility particle size spectrometers: harmonization of technical standards and data structure

to facilitate high quality long-term observations of atmospheric particle number size distributions. *Atmospheric measurement techniques*, 5(3):657–685.

[Zhou et al., 2020] Zhou, Y., Dada, L., Liu, Y., Fu, Y., Kangasluoma, J., Chan, T., Yan, C., Chu, B., Daellenbach, K. R., Bianchi, F., Kokkonen, T. V., Liu, Y., Kujansuu, J., Kerminen, V.-M., Petäjä, T., Wang, L., Jiang, J., and Kulmala, M. (2020). Variation of size-segregated particle number concentrations in wintertime beijing. *Atmospheric chemistry and physics*, 20(2):1201–1216.

[Zíková and Zdímal, 2013] Zíková, N. and Zdímal, V. (2013). Long-term measurement of aerosol number size distributions at rural background station kosetice. *Aerosol and air quality research*, 13(5):1464–1474.